Image Retrieval within Augmented Reality

Philip Manja

May 5, 2017

Technische Universität Dresden

Fakultät Informatik
Institut für Software und Multimediatechnik
Professur für Multimedia-Technologie

Master's Thesis

Image Retrieval within Augmented Reality

Philip Manja

1. Reviewer Prof. Raimund Dachselt
Fakultät Informatik
Technische Universität Dresden

2. Reviewer Dr. Annett Mitschick
Fakultät Informatik
Technische Universität Dresden

Supervisors Dr. Annett Mitschick and Wolfgang Büschel (M.Sc.)

May 5, 2017

Philip Manja
Image Retrieval within Augmented Reality
Master's Thesis, May 5, 2017
Reviewers: Prof. Raimund Dachselt and Dr. Annett Mitschick
Supervisors: Dr. Annett Mitschick and Wolfgang Büschel (M.Sc.)

Technische Universität Dresden
Professur für Multimedia-Technologie
Institut für Software und Multimediatechnik
Fakultät Informatik
Nöthnitzer Straße 46
01187 Dresden

Abstract

The present work investigates the potential of augmented reality for improving the image retrieval process. Design and usability challenges were identified for both fields of research in order to formulate design goals for the development of concepts. A taxonomy for image retrieval within augmented reality was elaborated based on research work and used to structure related work and basic ideas for interaction. Based on the taxonomy, application scenarios were formulated as further requirements for concepts. Using the basic interaction ideas and the requirements, two comprehensive concepts for image retrieval within augmented reality were elaborated. One of the concepts was implemented using a Microsoft HoloLens and evaluated in a user study. The study showed that the concept was rated generally positive by the users and provided insight in different spatial behavior and search strategies when practicing image retrieval in augmented reality.

Abstract (deutsch)

Die vorliegende Arbeit untersucht das Potenzial von Augmented Reality zur Verbesserung von Image Retrieval Prozessen. Herausforderungen in Design und Gebrauchstauglichkeit wurden für beide Forschungsbereiche dargelegt und genutzt, um Designziele für Konzepte zu entwerfen. Eine Taxonomie für Image Retrieval in Augmented Reality wurde basierend auf der Forschungsarbeit entworfen und eingesetzt, um verwandte Arbeiten und generelle Ideen für Interaktionsmöglichkeiten zu strukturieren. Basierend auf der Taxonomie wurden Anwendungsszenarien als weitere Anforderungen für Konzepte formuliert. Mit Hilfe der generellen Ideen und Anforderungen wurden zwei umfassende Konzepte für Image Retrieval in Augmented Reality ausgearbeitet. Eins der Konzepte wurde auf einer Microsoft HoloLens umgesetzt und in einer Nutzerstudie evaluiert. Die Studie zeigt, dass das Konzept grundsätzlich positiv aufgenommen wurde und bietet Erkenntnisse über unterschiedliches Verhalten im Raum und verschiedene Suchstrategien bei der Durchführung von Image Retrieval in der erweiterten Realität.

Contents

1 Introduction

The aim of the present work is to investigate the possibility of improving the image retrieval process with the help of augmented reality (AR). Being an entry to the thesis, section 1.1 motivates why such a connection between the two fields has potential. Section 1.2 is then giving an overview on the methodical approach and the resulting structure of the present work.

1.1 Motivation and Problem Statement

Connecting image retrieval with AR can provide advantages for the image retrieval process by making use of novel interaction techniques. On one side, head-mounted displays (HMDs) are currently on the rise with several devices introduced during the last years. Such devices offer a totally different way of interacting with data (see section 1.1.1).
On the other side, the importance of image retrieval is rising with the growing number of images. A new form of interaction provides the possibility for image retrieval to adopt to the changing needs of people (see section 1.1.2).
A conjunction of both field conforms to a general trend towards natural search interfaces and can provide a direct way of retrieving data when information needs are triggered by the surroundings (see section 1.1.3).

1.1.1 Augmented Reality and Head-Mounted Displays

Despite of the rapid development of technology and computer devices in the last decades there have only been minor changes in the way we interact with them. Desktop and laptop devices mostly still rely on the usage of a keyboard and a mouse, both of which have been introduced a long time ago. Only recently, with the introduction of powerful mobile devices, other interaction techniques like multi-touch were brought into our everyday life. This development shows that a fundamental change of interaction will probably rather likely be closely related to an introduction of new devices.

Mobile devices are not only an addition to the range of such devices. They rather depict an ongoing trend towards mobile and everywhere computing, with mobile usage of digital media overtaking desktops in 2013 [LL14]. A part of this development is the growing popularity of wearable devices such as smart watches and the recent introduction of HMD. While virtual reality (VR) headsets like Oculus Rift and

HTC Vive are mostly aimed at gaming and entertainment, AR outfits like Meta 2 or Microsoft HoloLens claim to change the way of everyday interaction with computer devices. This vision implies a shift of our computer activities to holographic devices similar to the one from stationary to mobile devices during the last decade. That shift would take interaction from surfaces into space. 3D interaction has been a topic in research for decades [Han97], however there are still no well-known and proven sets of gestures as there are for multi-touch surfaces. Therefore, in order to help establish such proven ways of interaction, the field of AR, especially the usage of HMDs, is an important recent field of research.

1.1.2 Image Retrieval

Humans are very visual beings. Whole social media platforms like YouTube, Instagram and Snapchat are built around the presentation and sharing of visual content. Nowadays, every owner of a modern mobile phone is able to shoot photos and videos in nearly every situation. The ever increasing amount of images, on the Internet as well as in every single personal collection, inherently makes the process of searching and retrieving more important than ever. Additionally, because different content is often represented by some kind of visual cue, research conducted in image retrieval can also be transfered to other domains.

The way we retrieve images today does not differ greatly from what it used to be 15 years ago. When searching for pictures on the Internet, keyword-based search is still the main kind of interaction. Browsing personal images is mostly done with pictures organized in folders. On the one hand, this might show that other forms of search are simply not sufficient enough. As Marti A. Hearst points out, search interfaces are used by nearly everyone, therefore having to rely on simple and understandable kinds of interaction [Hea09]. However there are some shortcomings of those classical ways of interaction. One of them is the semantic gap, which describes the discrepancy between the understanding of a picture from a user point of view and a system point of view.

This problem has been a part of research for many years [Sme+00]. One central approach to reduce the gap is to improve the system's understanding of the pictures, for example in the form of content-based image retrieval (CBIR). Today's advanced computer vision techniques are making such systems relatively reliable and robust. Hence another field of research has been growing recently: human computer information retrieval (HCIR). This approach concentrates on centering retrieval systems around the users, providing more powerful and satisfactory tools to achieve their goals. Altogether, image retrieval still has big potential as a field of research.

1.1.3 Image Retrieval within Augmented Reality

As there has been stated in section 1.1.2, one possible approach of reducing the semantic gap is to give the user new tools to retrieve images. Section 1.1.1 outlined the capability of HMDs and AR to provide new ways of interaction. Therefore the present thesis investigates how the potential of AR could be used to improve image retrieval processes.

According to Marti A. Hearst in [Hea11], a trend towards more natural search interfaces is observable: Users will

- speak rather than type,
- use full sentences rather than artificial keywords,
- watch video rather than read,
- use technology socially rather than alone,
- point with fingers rather than mice.

Using HMDs and AR, all these kinds of interaction can be realized. AR aligns virtual content in the real world, thereby offering the possibility to visualize and interact based on locations of the real world. In the case of image retrieval this means that querying, displaying and managing images can be relocated into the physical world. Because the need of information is often triggered by our surroundings, the direct connection between real objects and computer devices can allow for faster ways to start a retrieval process. Besides that, it allows to establish new forms of visualization and make use of objects and locations that are already familiar to users due to their usage in the everyday life.

Retrieval processes are often involving social situations in personal or work environments. Bringing content into the real world also means that collaboration can be made much easier compared to desktop applications. With HMDs, in contrast to large displays or tabletops, users can work collaboratively on the same data and still have different views of it. This means that a new kind of division of work is possible using HMDs for retrieval systems. Although collaboration scenarios are a reason for researching image retrieval within AR, the present work focuses on single users as a starting point.

As outlined by this section, numerous reasons indicate that the usage of AR and HMDs can be beneficial for image retrieval. The present work is therefore investigating both fields of research in order to develop a concept for image retrieval within AR that is then evaluated using a prototypic implementation.

1.2 Thesis Structure

The present thesis is divided into six chapters that are depicting the general workflow of the thesis as follows:

Chapter 2: This chapter depicts a foundation for both AR and image retrieval by defining the terms and investigating different design challenges. Based on these findings, a taxonomy for AR image retrieval systems is proposed in order to facilitate the examination of related work and the design of new concepts in chapter 3.

Chapter 3: This chapter investigates research work related to image retrieval within AR and proposes a number of basic interaction concepts which are classified using the taxonomy elaborated in chapter 2. Possible application scenarios and design goals are then formulated based on the design challenges identified in chapter 2. Finally, with the help of these concept requirements, two comprehensive concepts for image retrieval in AR are presented.

Chapter 4: This chapter describes the process and result of the implementation of one of the comprehensive concepts presented in chapter 3 using a Microsoft HoloLens. It also provides a documentation for users and developers and compares the final implementation with the original concept.

Chapter 5: This chapter explains the design of the user study that has been conducted using the prototype implemented in chapter 4 and presents the results that have been gathered using data logging and a questionnaire as well as free text remarks and observations.

Chapter 6: This chapter concludes the work by summing up the results and giving an outlook on future work in the field of image retrieval within AR.

2 Foundations of Image Retrieval and Augmented Reality

In order to successfully design concepts for image retrieval in AR, a comprehensive investigation of both domains is essential. In section 2.1 image retrieval is investigated while section 2.2 focuses on AR and HMDs. Both sections accumulate definitions, classifications and design challenges. In section 2.3, a taxonomy for image retrieval within AR is provided based on the previous sections. This taxonomy shapes the framework for investigating related work and proposing new concepts in chapter 3.

2.1 Foundations of Image Retrieval

This section examines the field of image retrieval by first defining the term itself (section 2.1.1) and then presenting possible classifications for image retrieval interfaces (section 2.1.2). In section 2.1.3, different design and usability challenges regarding image retrieval are pointed out.

2.1.1 Definition of Image Retrieval

One commonly used term in the field of image retrieval is CBIR. CBIR is a major part of research in image retrieval and concentrates on retrieving images based on their content and visual features. According to Datta et al., CBIR is "any technology that in principle helps to organize digital picture archives by their visual content" [Dat+08]. Given the fact that this present work is focusing on interaction techniques and visualization, no emphasis is placed on whether images are retrieved based on their content or based on meta data. Therefore in this work, the term *image retrieval* is used rather than the term CBIR.

Image retrieval is a branch of the more general information retrieval. Cambridge dictionary defines information retrieval as "the process of finding stored information on a computer" [1] while Oxford dictionary calls it "The tracing and recovery of specific information from stored data" [2]. The definition proposed by Merriam Webster reads

[1] `http://dictionary.cambridge.org/dictionary/english/information-retrieval`, accessed on December 10, 2016

[2] `https://en.oxforddictionaries.com/definition/information_retrieval`, accessed on December 10, 2016

"The techniques of storing and recovering and often disseminating recorded data especially through the use of a computerized system" [3].

For the present work, these definitions are merged to define image retrieval as follows:

> *Image retrieval are techniques and processes of tracing and recovering images from stored data through the use of a computerized system.*
>
> — **definition of image retrieval**
> based on definitions of information retrieval

2.1.2 Classification of Image Retrieval Systems

Based on the definition elaborated in section 2.1.1, four major parts of image retrieval sessions can be extracted:

- The *process* of tracing and recovering images.
- The *user* that is tracing and recovering images.
- The *system* with which the user is tracing and recovering images.
- The *data* that images are being traced and recovered from.

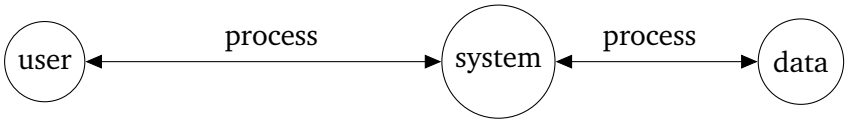


Fig. 2.1: Parts of image retrieval sessions

Through the process of tracing and recovering images, the image retrieval system is connected to both the user and the data (see figure 2.1). Therefore the system's properties depend on user and data as well. Hence it is possible to classify image retrieval systems based on the user, the data and the system itself.
Classifications from a system's point of view like the one proposed by Datta et al. differentiate between *text-based* and *content-based query processing* [Dat+08]. However due to the present work's focus on interaction and visualization, no further investigation is made into technical aspects of image retrieval systems. Therefore the following sections present different classifications only based on the user and the data.

[3] `https://www.merriam-webster.com/dictionary/information%20retrieval`, accessed on December 10, 2016

Classification based on the user

In an information retrieval session, user-related factors describe why and how users search for information. Although the user is not an integral part of the system, different intentions require different approaches. Therefore image retrieval applications can be classified based on the fact of which approaches they support (for example: “a browsing interface”).
Hollink et al. propose a classification of users based on three factors: *Domain*, *Expertise* and *Task* [Hol+04]. As the domain describes the data that is being searched on, it is investigated in the section about data. Although expertise is a property of the user, it is essentially just an instance of one level of expertise that a domain contains (see figure 2.2). Therefore, expertise is also explained in the data section. According to Hollink et al., the task of the user consists of a goal, a retrieval specification and a retrieval method, which are explained in this section [Hol+04].

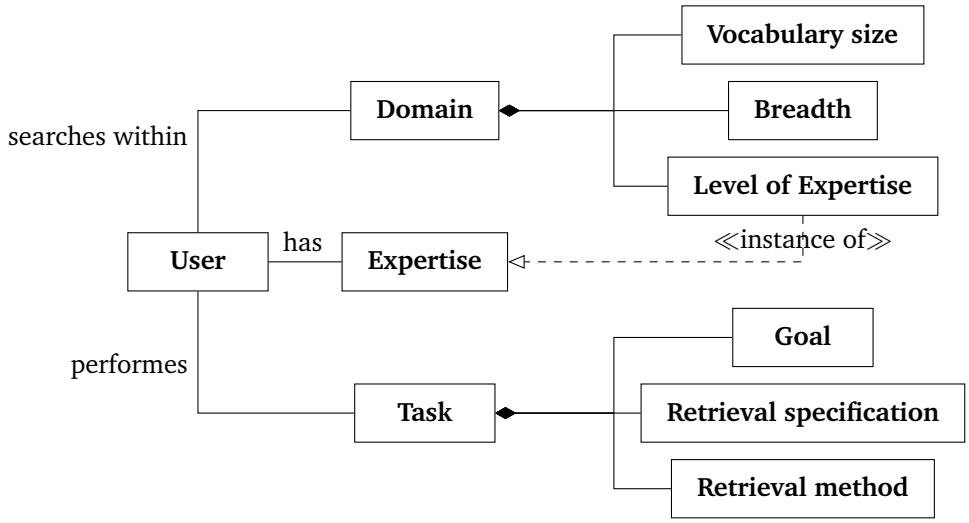


Fig. 2.2: Classification of users according to Hollink et al. [Hol+04]

Goal: According to Hollink et al., the user’s *goal* expresses why the search is being conducted in the first place [Hol+04]. Based on work of Raya Fidel [Fid97], the goal of the user can be placed on a continuum ranging from a *data pole* to an *object pole*. A goal near the data pole means that the user retrieves images to extract information from it, e.g. searching an image of a fruit to learn how it looks like. In contrast to that a goal near the object pole means that the user treats the retrieved images as objects, e.g. searching photos of fruits to decorate a room with prints of them. Different goals also have impacts on the search process: When searching for data, the user is aware of relevance criteria, which means it can be specified before retrieval. For object search, users mostly recognize the relevance when viewing the images.

Retrieval specification: *Retrieval specification* is coined by Hollink et al. as a "set of conditions the user expresses as input for the search" [Hol+04]. For different types of search they refer to the three categories of Smeulders et al.: *target search*, *category search* and *association search* [Sme+00]. Table 2.1 explains these categories and compares them to similar classifications by other authors: Cox et al. classify content-based image retrieval systems into three categories according to the aims of the user and Datta et al. are putting the focus on the clarity of the user's intent [Cox+00] [Dat+08].

Tab. 2.1: Different descriptions for retrieval specification

Cox et al. [Cox+00]	**Smeulders et al. [Sme+00]**	**Datta et al. [Dat+08]**
In *target search* the user wants to find one particular image.	In *target search* the user wants to find one particular image or another image of a particular object.	A *searcher* has a clear intent, using a short session with coherent searches to get her desired end-result.
In *category search* the user wants to retrieve images of a certain category.	*Category search* "aims at retrieving an arbitrary image representative of a specific class".	A *surfer* has a moderate intent, starting exploratory and increasing clarity with subsequent searches.
Open-ended search or *browsing* means that the user searches the database with a broad and non-specific goal.	*Association search* has no specific aim other than to find interesting things.	A *browser* has no intent at all, with sessions consisting of unrelated searches across multiple topics.

Retrieval method: According to Hollink et al. the *retrieval method* characterizes "the tactics the searcher uses to express the retrieval specifications" [Hol+04]. Possible retrieval methods named are browsing for association search, textual queries or multiple examples for category search and sketching for target search.
To describe these kind of retrieval methods, Datta et al. used the name *query modality* [Dat+08]. The modalities they differentiate are visualized in figure 2.3 and described as follows:

- *Keywords* are simple queries in the form of words.
- *Free-text* is a complex phrase, sentence, question, or story.
- *Image* means a query image is used to find similar ones.
- *Graphics* are hand-drawn or computer-generated pictures used as queries.
- *Composite* means using multiple modalities to query a system.

Fig. 2.3: Query modality according to Datta et al. [Dat+08]

Classification based on the data

The data that is being searched on is called *domain* by Hollink et al. [Hol+04]. They identify three aspects that characterize domains: The *breadth of the domain*, the *size of the vocabulary* and the *levels of expertise*. The size of the vocabulary is defined by the number of domain-specific terms and the ratio between these terms and general terms [Hol+04]. Breadth of the domain and levels of expertise are described as follows:

Breadth of the domain: According to Hollink et al., "the breadth of the domain is the variability of the images within the domain" [Hol+04].
Smeulders et al. define a spectrum based on the visual variability of the underlying data (see figure 2.4) [Sme+00]. On the one side of the spectrum, a *narrow domain* has a limited content variability with homogeneous semantics and specific applications. On the other side, a *broad domain* has high variability, heterogeneous semantics and generic applications.

Fig. 2.4: Image Domain according to Smeulders et al. [Sme+00]

A more specific classification named *data scope* has been proposed by Datta et al. [Dat+08]. This classification is mainly based around a user's point of view on the data, describing some concrete kinds of domains:

- A *personal collection* is largely homogeneous and small in size, accessed mostly by its owner.
- A *domain-specific* collection is homogeneous and may be large in size, accessed by controlled users with specific objectives.
- An *enterprise* collection is heterogeneous and may be accessed by users within an organization in a uniform or nonuniform way.
- *Archives* are homogeneous and large in size, containing structured or semi-structured data affiliated with certain topics, accessible by most people on the Internet.
- *Web* image collections are heterogeneous and massive in size, semi-structured and accessible by virtually anyone.

Level of Expertise: According to Hollink et al. every domain has certain levels of expertise, of which the user can have one of them [Hol+04]. Levels of expertise are characterized by differences in the amout of knowledge between experts and non-experts, the occurence of intermediate levels, the effort needed to become an expert and the basic level of knowledge about the domain [Hol+04].

Summary of Image Retrieval Classifications

This section has shown different ways of classifying image retrieval systems. Based on the definition of image retrieval, three parts for classification have been identified: user, system and data. While classifications based on the system have not been investigated, several possibilities to classify image retrieval based on user and data have been presented.

2.1.3 Design and Usability in Image Retrieval

When designing image retrieval interfaces, several different levels of interface design should be taken into account:

1. General usability guidelines
2. Search interface design
3. Image retrieval challenges

Recapitulating general usability guidelines is not within the scope of the present work. To coarsely cover the second item, the information seeking process and general search user interface guidelines based on Marti Hearst are summarized in the following sections. The last section then investigates concrete design challenges of image retrieval interfaces in a more detailed way.

Information Seeking Process

When designing an interface for search it is essential to know how people are generally seeking for information. The different models that describe this process are summarized based on the work of Marti Hearst [Hea09].

Standard model: The standard model consists of a sequence of steps that are usually iterated in the process of information retrieval. A typical description of the standard model contains steps like *identifying the information need*, *expressing the query to a system*, *reviewing the results* and *reformulating the problem* [Hea09].

Cognitive models: Based on work of Norman [Nor88], a cognitive approach for information retrieval can be elaborated: With a certain search goal in mind, users decide on query actions based on their mental model of the retrieval system. The result is then evaluated and the process begins anew. The gap between the intention and the outcome is called gulf of execution, while determining if the goal has been met is called gulf of evaluation [Nor88].

Dynamic model: While standard and cognitive models are based on a static information need, studies have shown that needs change when interacting with the

system [Hea09]. Search results for one goal will likely trigger new goals within the same context but in some new direction. The main value of the search process is the learning and acquisition of information that is conducted during the process. According to the so called berry-picking model, the information need constantly shifts and is not satisfied by a single final set but series of selections [Hea09].

In stages: To describe information seeking with a complex information need over extended periods of time, different stages regarding knowledge and attitude of the user have been elaborated by C.C. Kuhlthau in 1991 [Kuh91]:

1. *Initiation:* seeking background information and recognizing tasks with general, vague thoughts and uncertainty
2. *Selection:* optimistically identifying tasks
3. *Exploration:* seeking relevant information and investigating tasks with clearer thoughts, confusion, frustration and doubt
4. *Formulation:* formulate tasks and feel clarity
5. *Collection:* seeking relevant or focused information and gathering tasks with increased interest, sense of direction and confidence
6. *Presentation:* completing tasks with clear and focused thoughts, feeling relief and satisfaction or disappointment

As strategic process: Different kinds of models that describe information seeking as strategic process have been proposed [Hea09]:

- *Sequences of Tactics*: Different types of tactics like query reformulation, information structure and monitoring are executed in a sequence changed by triggers and ended by stop conditions [Hea09].
- *Information Foraging*: Like in food-finding, cost-benefit analyses between expected information (information scent) and effort of discovering this information are applied [Hea09].
- *Browsing and Searching:* Retrieval is done by selecting pre-defined categories (browsing) or by issuing search queries in order to get ad hoc collections (searching) [Hea09].
- *Orienteering:* Searchers break down complex problems into simple steps, often starting with imprecise approximations to get into the right part of information space and concluding with local operations to meet the goal [Hea09].

Sensemaking: According to Marti Hearst "sensemaking refers to an iterative process of formulating a conceptual representation from of a large volume of information" and is often applied to complex tasks that make a kind of information organization necessary [Hea09]. It divides the process of information seeking into two main components: *information retrieval through search* and *analysis and synthesis of results* [Hea09].

General Search Interface Design Guidelines

Based on general user interfaces guidelines, Marti Hearst has elaborated a number of guidelines for search user interfaces [Hea09]. Table 2.2 summarizes the proposed guidelines and best practices.

Tab. 2.2: Guidelines for search interfaces according to Marti Hearst [Hea09]

Offer efficient and informative feedback • Show result immediately • Show relation between result and query • Allow result sorting by various criteria • Show query suggestions • Indicate relevance with caution • Support rapid response
Balance user control with automated actions • Ordering of result • Transformation of queries
Reduce short-term memory load • Suggest action that happens when the user starts the search • Support simple history mechanisms and access to previous queries • Integrate navigation & search (provide categories & hierarchical metadata)
Provide shortcuts • Alternative interface mechanisms for practiced users • Hints about next steps of interaction
Reduce errors • Avoid empty result sets • Consider that different users express the same desires differently

Image Retrieval Challenges

Smeulders et al. divide the general process of retrieving images into *Query Specification*, *Query Space Display* and *Interaction with Query Space* [Sme+00]. These three steps each have their own challenges, which are investigated in this section: The specification of queries needs a *description of an image* in the user's mind to pose a *query with a certain target* and overcome the *semantic gap*. The query space display involves certain *requirements* and different *kinds of visualizations*, i.a. based on *similarity* or *episodic memory*. Query space interaction can be realized by providing *relevance feedback* or using *workspaces*.

Image descriptions: One approach to find out which kind of queries users should be able to specify is to study their way of describing images in general. Hollink et al. conducted such a user study for category search where users had to describe an imaginary image [Hol+04]. They introduced three categories of possible descriptions:

- *conceptual descriptions* that describe the semantical content of the image.
- *perceptual descriptions* that describe visual properties like composition and color of images.
- *non-visual metadata* that is not apparent on the image itself.

They found out that 87% of all formulated descriptions were conceptual, with 70% of them concerning objects in the image while 30% described the scene. While the perceptual descriptions concerned scene and object roughly the same, over two thirds of them were describing the composition and the color of the content. However they also pointed out that results will likely be very different in target search when users are able to describe non-visual and perceptual properties of the concrete image they have in mind.

Query targets: Queries are a core component of any retrieval system and therefore especially important for a successful design of image retrieval interfaces. To understand the nature of image queries, Smeulders et al. have proposed a classification into two broad categories [Sme+00]:

- An *exact query* evokes a set of images fulfilling certain criteria.
- An *approximate query* evokes a ranking based on similarity to the query criteria.

For each of the two categories mentioned above, Smeulders et al. define three subclasses characterizing the target of the query [Sme+00]:

- *spatial content of the image*: the query is targeting visual or semantic features within the image
- *global image information*: the query is targeting properties of the image with no spatial basis
- *groups of images*: the query is targeting categorical properties belonging to groups of images

Approximate queries are mostly specified by examples like sketches (spatial content of the image), images (global image information) or groups of images. Exact queries are specified by predicates like “sun over water” (spatial content of the image), “mostly blue” (global image information) or “Location = Africa” (groups of images) [Sme+00].

Semantic gap: As Hollink et al. have shown in their study, users mostly use conceptual descriptions of images when searching for a certain category [Hol+04]. To understand these descriptions, a system has to analyze the images and extract these kind of concepts. This discrepancy between the meaning of a picture to the user and the understanding of a picture by the system is called the *semantic gap*. According to Smeulders et al. in [Sme+00], “the semantic gap is the lack of coincidence between the information that one can extract from the visual data and

the interpretation that the same data has for a user in a given information". In other words, the problem is "that the user seeks semantic similarity but the database can only provide similarity by data processing" [Sme+00].
Several approaches have been made to reduce the semantic gap. Generally said, the semantic gap can be narrowed down in two ways: Firstly, by making the system understand the image as good as possible in order to make assumptions what the user's perceived meaning of the picture is. This is the approach of CBIR, where the retrieval is based on the content of the image in order to give the system the possibility to analyze an image in a similar way to a user. Another possibility however is to give the user specific tools to express her wishes in a way the system understands best. This user-centered approach falls into the category of HCIR.

Requirements for visualizations: Due to the visual nature of human perception and the visual character of images, visualization is a central part of displaying the retrieved images to the user. Nguyen and Worring identified three major requirements that every visualization of large image databases should fulfill [NW08]:

1. *Overview*: Give a faithful overview of the image distribution in the collection
2. *Structure preservation*: Preserve relations between images
3. *Visibility*: Present each image in a way that users can understand the content

One possible approach to preserve structure and give an overview is a graph based search as proposed by Worring et al. [Wor+07]. A widespread approach for the arrangement of images is similarity based visualization.

Types of visualizations: To distinguish different design approaches for the visualization of images, Datta et al. provide a categorization as follows [Dat+08]:

- *Relevance-Ordered* means that results are ordered by a numeric measure of relevance to the query.
- *Time-Ordered* means that results are shown in chronological order.
- *Clustered* means that results are clustered by their metadata or visual content.
- *Hierarchical* means that results can be arranged in tree order based on their metadata.
- *Composite* means that two or more forms of visualization are mixed.

Similarity-based visualization: Rodden et al. compared a visualization based on captions with one based on visual similarity and found out that the visualization based on captions is useful to break down the data according to meanings [Rod+01]. They found out several things: The usefulness depends on the quality of the captions. Also, users might need labels to understand the structure. Displaying images based on visual similarity helps to divide the data into more simple categories but might prevent individual images from sticking out. Therefore when browsing pictures without particular requirements, random arrangements might be more useful. For

some users, accessing different arrangements of pictures proved to be useful while for others it was a distraction [Rod+01].
In a database annotation scenario, using an objective user model, the proposed similarity based visualization of Nguyen and Worring reduced the total annotation effort by up to 16 times compared to a classical one [NW08]. Schoeffmann and Ahlstrom compared a color-based sorted storyboard visualization to an unsorted one in a user study and found out that the sorted one was 20% faster and preferred by ten of twelve users [SA11]. Strong and Gong have shown that their similarity-based image browsing technique using intuitive zooming, panning and image resizing can reduce time for users to find a desired image [SG11].

Episodic memory: As shown before, Hollink et al. have found out that users mainly describe images based on concepts rather than visual features in category search [Hol+04]. One possible approach of visualization of personal collections is to cluster images by time and location [Che+06]. These two factors were chosen by Chen et al. because they are related to human episodic memory [Che+05]. In a user study, they have shown that the total system searching time as well as the questionnaire responses were significantly better for their visualization in comparison to standard image browsers [Che+06].

Relevance feedback: In research, interactive information retrieval often refers to the iterative process of refining queries. *Relevance feedback* has been introduced to image retrieval by Rui et al. in order to give the user a tool to refine her queries by telling the system which results were relevant to the query [Rui+98]. According to Datta et al., relevance feedback is "a query modification technique which attempts to capture the user's precise needs through iterative feedback and query refinement" [Dat+08]. Different advancements for relevance feedback have been identified and collated by Datta et al. [Dat+08]:

- *Learning-based* approaches modify the feature set or similarity measure.
- *Feedback specification* approaches use alternative ways like semantic labels or groups of images to let the user communicate feedback.
- *User-driven* approaches provide the user with cues to improve query formulation and model the users mental image.
- *Probabilistic* models represent the user's goals by a distribution.
- *Region-based* approaches consider different region of interests (ROIs) for the user.

Workspaces: Urban and Jose have shown that a workspace can be beneficial when retrieving images [UJ06]. Working with a workspace helped people to conceptualize and diversify the task better and increase the effectiveness, which means that less queries were issued to find a larger selection of images. Workspaces allow people to leave footprints and follow up multiple trains of thought. The workspace was

also especially useful for creating groups and giving feedback as well as receiving recommendations [UJ06].
Although it was more difficult to use and the cognitive effort required to solve a task was higher, the perceived usefulness of the workspace increased with the complexity of the tasks. All in all, two thirds of the users preferred a workspace environment over a classical retrieval interface. The latter however was suited better for tasks that required selection of a large number of images for a very specific topic [UJ06]. More detailed results can be found in [UJ07].

Summary of Image Retrieval Design Challenges

To give an overview on different challenges when designing image retrieval interfaces, general information seeking models and search user interfaces guidelines have been gathered as well as specific image retrieval design challenges.
Standard, cognitive and dynamic models to describe the information seeking process have been adapted from Marti Hearst as well as the process in stages, as strategy and sensemaking [Hea09]. General search user interface guidelines have been summarized based on work of Marti Hearst [Hea09]. Finally different aspects of importance during the image retrieval process, namely image descriptions and queries, the semantic gap, different result visualizations, relevance feedback and workspaces have been outlined.

2.2 Foundations of Augmented Reality

Beneath the application domain of image retrieval, AR as underlying technology has to be investigated as well in order to design interaction concepts. Therefore this covers the field of AR by giving a definition of the term in section 2.2.1 and pointing out design challenges in section 2.2.2. Due to the focus on interaction and visualization, technical aspects of AR are not part of the present work.

2.2.1 Definition of Augmented Reality

The most wide-spread definition of AR has been proposed by Ronald Azuma in 1997 [Azu97]. According to him, an AR-system

1. combines real and virtual content
2. is interactive in real time
3. is registered in 3D

Although this definition is already two decades old, it is still used to characterize modern AR interfaces. Therefore this definition is used as well in the present work.

2.2.2 Augmented Reality Design and Usability

When designing user interfaces, different aspects have to be taken into consideration. Since the present work is investigating how AR based on HMDs can act as a provider of new kinds of interaction for image retrieval, this section concentrates on interaction design and different modalities.
Table 2.3 shows classifications of AR user interfaces according to Billinghurst et al. and Schmalstieg and Höllerer [Bil+15][SH16]. Although they partly use different terms, five common kinds of interfaces can be identified.

Tab. 2.3: Possible classifications of AR interfaces

Description	**Billinghurst et al. [Bil+15]**	**Schmalstieg and Höllerer [SH16]**
The user utilizes AR as a window to browse an information space.	information browsers	augmented browsing
Traditional 3D user interface techniques are used to interact with virtual objects. This may contain physics and haptics.	3D user interfaces	physically based interfaces, haptic interaction
Virtual information is overlaid on physical objects that are used for interacting with the application.	tangible augmented reality	tangible interfaces
Body motion and gestures are used for interacting with the application.	natural user interfaces	body tracking, gestures and touch
Multiple modalities like speech and gesture are combined to interact with the application.	multimodal interfaces	multimodal interaction

For image retrieval in AR, each of the proposed categories would be feasible: Retrieving results makes a kind of *information browsing* necessary. *3D interface techniques* could be used to display and manipulate results in space. *Tangibles* could be used to specify queries. *Natural user interface techniques* are a general trend in search interfaces, which also involves the combination of *different modalities*.
Nevertheless it is out of scope for this work to investigate design and usability challenges of all these types of interfaces in detail. Instead, challenges that all AR applications have in common are outlined in this section. Based on the definition of AR mentioned in section 2.2.1, Billinghurst et al. name three major parts that are crucial for designing AR interfaces [Bil+15]: *Real objects*, *virtual objects* and an *interaction metaphor* that connects both.
Based on this classification, the challenges of AR systems are investigated by examining the real world, the virtual world and the interaction metaphors as part of AR user interfaces.

Real world

The definition of AR requires the registration of content in 3D, which forms the connection between the real and virtual world. According to Schmalstieg and Höllerer, the real world can thereby serve as anything from a plain backdrop to a central modality for interacting with the system [SH16]. This section investigates the role of the real world for input and output.

Output in the real world: To use the real world as output for AR, virtual content has to be registered and displayed. According to Schmalstieg and Höllerer, registering content can be done in two ways: Locally and globally [SH16]. Local registration means that the coordinate system of the virtual content aligns to a movable object. Global registration in contrast means that the coordinate system is not connected to a certain object but rather to a location.
Although the definition of AR requires virtual content to be registered with the real world in 3D, it does not have to be semantically related to the environment. For content that does so, White and Feiner use the term *situated visualization* [WF09]. According to Schmalstieg and Höllerer, placing virtual objects into free space is possible, but aligning them with a real object makes them intuitively understandable and easier to interpret. They name different categories of real objects that can be used to provide a reference for augmentation [SH16]:

- *Vertical surfaces* simulate wall-mounted objects like pictures.
- *Horizontal surfaces* such as desktops that can be augmented with 2D content or used as supporting surfaces for 3D content.
- *Body-referenced* augmentations on the user's own body.
  - *Head-referenced* displays remain in the user's view and are useful for the continuous display of status information.
  - *Torso-referenced* displays that are either directly attached to the torso or extend the body into space.
  - *Hand-referenced* displays can resemble real objects held in the hand, making them agilely movable and manipulatable with the other hand.
  - *Arm-referenced* displays have similar properties like hand-referenced displays. They are not as natural, but they can be used when the user's hands are occupied with real objects.

Schmalstieg and Höllerer name three advantages for using body-referenced augmentation [SH16]:

1. The user's body is always available.
2. Relying on body parts avoids any instrumentation of the user.
3. Human proprioception allows intuitive interaction with the augmentation.

Input of the real world: Using objects in the real world as instruments for input is called *tangible interaction*. According to Preim and Dachselt, "tangible user interfaces augment the real, physical world by attaching digital information to everyday material objects and surroundings" [PD15]. This definition clearly shows the affinity between tangible interaction and AR. Tangible AR has been proposed in the first place by Kato et al. in 2000 in order to overcome shortcomings of tangible user interfaces by combining them with the enhanced display possibilities of AR [Kat+00].

Tab. 2.4: Operations of generic tangibles in AR according to Schmalstieg & Höllerer [SH16]

Operation	**Effect**
translation and rotation of single tangibles	manipulate objects, modify parameters
spatial relationship of multiple tangibles	arrangement of objects
distance between two tangibles	scalar value, e.g. association between two objects
removing or covering tangible	triggering command, memorize location
shaking, turning, circular motion, tilting, pushing tangibles	gestural input

According to Schmalstieg and Höllerer, tangibles in AR are often used for focus+context interactions and can either have generic or distinct shapes. While generic shapes offer a rich variety of input operations (see table 2.4), distinct shapes can provide immediate recognition due to their well-known affordances. This includes tools like paddles or flashlights as well as containers like tablets, books or boxes. In a similar manner, Billinghurst et al. divide tangible interfaces into space multiplexed interfaces, where each physical tool is dedicated for one function, and time multiplexed, where tangibles have many functions and purposes [Bil+15]. Lee et al. propose a spectrum for the relationship between the tangible prop and the interaction object that ranges from abstract (generic shape) over metaphoric (class-related shape) to concrete (object-related shape) [Lee+07]. Additionally, they divide the mapping into direct and indirect as well as static and dynamic [Lee+07]. According to Lee et al., tasks in tangible AR interfaces can be classified into three groups [Lee+07]: The first group is viewpoint control (*navigation*) that is usually defined by the technology that is used. The second concerns selection, release and 3D *spatial manipulation* of objects and the third are *event generation and system commands* that are triggered based on location, pose and gestures. Lee et al. postulate two major guidelines for tangible AR interfaces [Lee+07]:

1. Use metaphors like physics, interactions and objects from every day life.
2. Take advantage from parallel activities. Use both hands and multiple people, objects or interfaces.

Virtual world

Comparably to the varying complexity of the meaning of real objects for AR interfaces, virtual objects can be anything from a plain label to a complex multi-view interface. To address the different challenges, this section depicts general challenges of AR visualization as well as those of virtual user interfaces and multi-view interfaces.

Depth perception: To convincingly add virtual components into the real world the user has to perceive depth cues correctly in order to avoid cognitive problems [Kal+11]. Although a perfect alignment should prevent such problems, primary depth cues derived from real world features might not be replicated perfectly in 3D visualizations [PD15]. Therefore artistic depth cues like silhouettes, feature lines or hatching can be used in visualization in order to support perception of depth [PD15].

Data overload: According to Preim and Dachselt, 3D user interfaces should feature virtual content that is displayed only when relevant and a plausible way [PD15]. They state that AR systems are especially prone to data overload because they combine real and virtual objects. As Schmalstieg and Höllerer point out, an augmentation with HMDs can become disturbing when there is no possibility to switch it off [SH16]. To overcome this problem, the use of data transformation (filtering) in combination with a visual mapping and a transformation of the view (layout) is proposed as a possible approach to reduce visual cluttering [SH16].

Temporal coherence: One important aspect of AR is that the context can change over time or from place to place. Two main problems might arise because of that [Kal+11]:

- visual objects hiding important real-world objects
- shapes and colors of virtual objects blending in

In other words, virtual and real information can get lost during the process of visualization in AR. Therefore visualizations should be able to adapt to these changes in order to preserve their visual information [SH16].

Virtual user interfaces: Whenever complex interactions are needed in AR, virtual user interfaces can be used. Besides the opportunity to add complex interfaces, the possibility to reuse familiar solutions from desktop or mobile devices is advantageous as well [SH16]. Additionally, virtual interfaces can also be used to augment tangibles, giving them a tool character and using physical properties of the tangible as constraints or affordances [SH16].

Multi-view interfaces: The nature of AR allows to employ a variety of different multiple views (see table 2.5). In contrast to these coordinated views that are are

implicitly synchronized, cross-view interaction enables explicit synchronization by users [SH16].

Tab. 2.5: Multi-view interfaces according to Schmalstieg and Höllerer [SH16]

Name *Usage*	**View-points**	**Displays**	**Coordinate Systems**
Focus + Context	one	multiple	one
complementary information (2D/3D, high-res/low-res, mobile/stationary)			
Shared Space	multiple	one per viewpoint	one
collaboration			
Focus + Context Shared Space	multiple	multiple per viewpoint	one
overview map and multiple individual views			
Multiple Locales	multiple	multiple per viewpoint	multiple
arranging purely virtual objects, augmenting generic tangibles, combining egocentric & exocentric viewpoints			

Interaction Metaphors

In order to suggest which interaction metaphor has been employed, affordances on virtual and real objects are especially important [Bil+15]. Preim and Dachselt state that 3D interactions should have advantages in comparison to simple user interfaces (UIs) or non computer interfaces [PD15]. Different interaction techniques have been proposed for 3D interaction. They can be divided into three broad categories as defined by Bowman: navigation, selection and manipulation [Bow+04]. Preim and Dachselt name exploration, orientation and navigation as major 3D interaction techniques [PD15]. For AR applications, navigation and orientation are mostly done by moving in the real world. Therefore the three categories selection, manipulation and exploration are in the focus in this section.

Selection: For the selection of objects, one can distinguish between direct picking and indirect capturing [PD15]. For *direct selection*, hand tapping or a pointer metaphor is commonly used to select a single object [PD15]. A snapping algorithm can be employed to aid targeting the right object [PD15]. *Indirect selection* often employs a flashlight metaphor to capture objects inside a certain "cone of light" in front of the user [PD15]. When multiple objects come into consideration, the ambiguity can be resolved by inputting the name of the object or by choosing a representative of it [PD15].

Manipulation: The manipulation of objects can be divided into placement, scaling and rotation of objects [PD15]. The *placement* of objects is often realized through a

dragging interaction and can be supported by snapping, reference objects and carrier surfaces [PD15]. For *scaling and rotation*, well-known gestures from touch devices can be utilized and supported by clutching mechanisms [PD15]. Complex metaphors used for manipulation of objects are virtual model kits, 3D puzzle metaphors or virtual workbenches.

Exploration: For the exploration of visual data, the visual information-seeking mantra of overview first, then zoom and filter and details on demand should be taken into consideration [SH16]. Beneath overview techniques like maps, different filtering and detail views have been proposed for AR.
Schmalstieg and Höllerer divide information filtering into knowledge based and spatial filters [SH16]. Magic Lenses that can employ semantic zoom are often used with tangibles and can be seen as a combination of knowledge-based and spatial filters [SH16]. Other exploration metaphors like cutaways or x-ray visualizations and ghosting use space on the augmented object or make use of additional space, like explosion diagrams and space distortion [SH16].

Summary of Augmented Reality Design Challenges

In section 2.2.2 design and usability of AR systems have been examined. Different types of interfaces like tangible and natural user interfaces, information browsers and 3D user interfaces have been identified and linked to image retrieval. The role of the real world as output for AR has been inspected through the means of augmentation registration and placement while the input side has been covered by kinds and tasks of tangibles. Challenges of the visualization of virtual objects like depth perception, data overload and temporal coherence have been investigated as well as virtual and multi-view user interfaces. Lastly, different interaction metaphors have been outlined for selection, manipulation and exploration of objects in AR.

2.3 Taxonomy for Image Retrieval within Augmented Reality

This section builds upon the previous sections of this chapter and takes both fields of research into account to propose a classification in order to structure related work and form a framework for scenarios and concepts in chapter 3. More specifically, the taxonomy assures diversity in three important steps: Firstly, reviewing various related work that potentially contributes to the design of AR image retrieval applications. Secondly, defining different scenarios in which the use of an AR image retrieval system could be beneficial. Thirdly, proposing a variety of AR interaction concepts for image retrieval based on the reviewed work.

In section 2.1.2, four parts of image retrieval sessions have been identified: user, process, system and data. Due to the focus on interaction and visualization of the present work, the process of retrieving images is a main part of the taxonomy (see section 2.3.2) while the other parts of the session are grouped as session parameters (see section 2.3.1).

2.3.1 Session Parameters

Characteristics of user and data are suited well to describe an image retrieval session in a static context (see section 2.1.2). However in practice, especially for interaction with AR systems, the context has to be taken into consideration as well when describing a session. Therefore *user*, *data* and *context* are defined as session parameters and described in the following sections. As the session parameters are used primarily to describe and compare different scenarios for image retrieval, they depict rather qualitative measures.

User

In dependence on the classifications gathered in section 2.1.2, the following characteristics are used to describe users of AR image retrieval applications: *Goal*, *strategy* and *expertise*.

Goal: Adapted from Raya Fidel, the user's goal signifies whether she aims to use the retrieved images as a source of data or as objects [Fid97]. For the present work, the following characteristics are used:

- *Data-driven:* The user searches for certain data on the pictures.
- *In-between:* The user searches for images as source of data and object likewise.
- *Object-driven:* The user searches for images to handle them as objects.

Strategy: The term strategy represents retrieval specification as proposed by Hollink et al. and uses characteristics based on Cox et al., Smeulders et al. and Datta et al. (see table 2.1) as follows [Hol+04]:

- *Browsing:* The user searches for images with no particular target in mind.
- *Target search:* The user searches for one or more particular images.
- *Category search:* The user searches for one or more images of a particular category.

Thereby different aspects of the information seeking process can be taken into consideration: During the process, the user can be triggered to change the strategy. For example, when browsing, the user identifies a desire for a certain category of pictures and switches from browsing to category search. The process will end when

certain stop criteria are met. For a browsing session, these criteria comply with the berry-picking model and are reached after enough images have been gathered during the browsing process. For target search the retrieval of certain images will lead to a stoppage of the process.

Expertise: The user's expertise is a value along the levels of expertise that the domain provides. In order to roughly compare different scenarios with varying expertise, three simple levels of expertise are used in this work:

- *Low:* Little or no prior knowledge is needed to fulfill the image retrieval task.
- *Medium:* Some prior knowledge is needed to fulfill the image retrieval task.
- *High:* A high amount of prior knowledge is needed to fulfill the image retrieval task.

Data

This part of the classification is based on the data scope as proposed by Datta et al. [Dat+08]. The different values that Datta et al. introduce for the data-scope are described using four different characteristics: How homogeneous the data is, how large the data is in size, which people are able to access the data and how structured the data is.
All of these factors are used in the present classification as well: The aspect of the access is combined with the levels of expertise as described by Hollink et al. and is called *expertise breadth* [Hol+04]. Homogeneity is represented by the breadth of the domain as proposed by Smeulders et al. and called *domain breadth* [Sme+00]. Finally, *size* and *structure* are used as own characteristics as well.

Size: The size of the domain signifies how many images are contained in the data that is being searched on and is based on categories proposed by Datta et al. [Dat+08]:

- Collections *large* in size like the Web and archives.
- Collections with *medium* size like domain-specific and enterprise data.
- Collections *small* in size like personal collections.

Structure: This parameter indicates whether or not the image database contains some underlying data that can be utilized in the search process.

- *Structured* data contains rich metadata and hierarchies that facilitate browsing.
- *Semi-structured* data contains some metadata or hierarchy.
- *Non-structured* data contains no metadata or hierarchy at all.

Domain Breadth: The breadth of the domain is based on the proposition of Smeulders et al. [Sme+00]. To compare different breadths of domains, three simple gradations are used:

- *Broad* domains have high content variability with heterogeneous semantics and generic applications.
- *Medium* domains have some variability of image content and semantics and different possible applications.
- *Narrow* domains have limited content variability with homogeneous semantics and specific applications.

Expertise Breadth: Based on the assumptions of Hollink et al., the following breadth of expertise is used in this work [Hol+04]:

- *Broad:* Many intermediate levels between experts and non-experts exist, a big effort is needed to become an expert.
- *Medium:* Some intermediate levels between experts and non-experts exist, a considerable effort is needed to become an expert.
- *Narrow:* Few intermediate levels between experts and non-experts exist, only little effort is needed to become an expert.

Context

The aim of this part of the taxonomy is to propose some simple means of categorization to facilitate the design and evaluation of scenarios and concepts. Regarding image retrieval in AR, two important factors emerge based on the definition of the two topics: AR aligns real and virtual objects in 3D, which means that *location* is an important factor. Image retrieval is defined as a process, which means that the *timespan* is an important factor.

Location: Within the image retrieval process in AR, the location of the user can either change or stay the same. A change of location might have an impact on the user's task. Therefore, three different levels are elaborated:

- *Static:* The location does not change.
- *Dynamic:* The location is changing and has an effect on the retrieval task.
- *Location-independent:* The location may change but has no effect on the retrieval task.

Timespan: Based on differing models of the information seeking process (see section 2.1.3), different timespans are confined in this work and presented in table 2.6.

Tab. 2.6: Different timespans in image retrieval processes

Name	Description	Models
Complex stages	The user traverses different stages of complex retrieval tasks and sensemaking.	Information seeking in stages Sensemaking
Strategic process	The user exploits different dynamic retrieval possibilities and employs changing strategies.	Dynamic model Strategic process
Simple session	The user undertakes a session using multiple queries to achieve a certain goal.	Standard model Cognitive model
Single query	The user issues only one query.	-

2.3.2 Interaction Process

The image retrieval process can be divided into three parts of interaction as proposed by Smeulders et al.: *Query Specification*, *Query Space Display* and *Interaction with Query Space* [Sme+00]. However these three parts are renamed in the present work to emphasize the focus on AR and natural interaction.
Based on the trend towards natural search interfaces identified by Marti Hearst, the first step is called *Natural Query Specification* [Hea11]. The second step is named *Situated Result Visualization* based on the term used by White and Feiner [WF09]. The third step incorporates 3D Interaction into Query Space Interaction and is therefore called *3D Result Interaction*. Each of the three steps is described in detail in the following sections.

Natural Query Specification

A query can be seen as a kind of message that the user emits to the system. Based on the well-known communication model of Shannon and Weaver, the four characteristics that are used to describe natural query specification are derived in table 2.7 [Sha48].

Tab. 2.7: Query characteristics derived from the Shannon-Weaver model of communication [Sha48]

Model	Query-related Question	Proposed Property
Message	What is it describing?	The target of the query.
Encoding	How is it looking like?	The modality of the query.
Medium	How is it delivered?	The interaction modality.
Noise	How hard is the delivery?	The effort of the query specification.

Query Target: Smeulders et al. have proposed three levels that an image retrieval query can target [Sme+00]. For the present work, these three levels are defined as follows:

1. *Spatial Composition:* The arrangement of conceptual or perceptual features on the image.
2. *Image Properties:* Global characteristics of the image like colors and brightness.
3. *Group Semantics:* Non-visual metadata and semantics that arrange images into groups.

Query Modality: A classification based on query modality has been proposed by Datta et al. (see section 2.1.2) [Dat+08]. This classification is adapted as follows to fit an AR application:
Instead of keywords, which depict a classical and constrained query modality, the broader term *property descriptions* is used. This term incorporates simple keywords but also more advanced retrieval techniques like faceted search.
Images and graphics are essentially different sections on one continuum of *visual examples*. A visual example used as a query can illustrate anything from a rough sketch over an arrangement of icons and patches of textures to an actual photograph. Therefore, query modality can take on the following occurrences in this work:

- *Free Text:* A query with natural language description of the desired images.
- *Visual Examples:* A query with visual representations of the desired images.
- *Property Descriptions:* A query that has neither visual nor free text form, for example key-value pairs, faceted search parameters or semantic categories.

Interaction Modality: The interaction modality describes the kind of interaction the user can employ to put her query across to the system and is derived by the different possible interactions used for AR systems. The modality used for the query (i.e. how the query is encoded) can differ from the modality of the interaction (i.e. how the query is delivered): For example a free text query can be delivered to the system using voice or typing. Therefore when focusing on interaction concepts it is necessary to distinguish between both forms. For the current work, the following interaction modalities are differed: *Voice Input*, *Freehand Gestures*, *Tangible Input*, *Body Motion* and *Gaze Input*.

Effort: The effort to deliver the query signifies the amount of explicit interaction that the user has to employ in order to send a query to the system.

- *Automatic:* The query is specified and sent without explicit help of the user.
- *Semi-automatic:* The query is specified by the user but sent automatically.
- *Manual:* The user specifies the query and submits it whenever she wants to.

Situated Result Visualization

Situated result visualization consist, as the name suggests, of a visualization of the retrieved result that is placed somewhere in space. A situated visualization is connected to the real world in two ways: The location itself and the meaning of the location [WF09]. For image retrieval, the arrangement of images in the visualization is of importance. Therefore the following parameters can be identified for situated result visualization:

1. *Reference*: Where is the visualization placed?
2. *Result Space*: How is it semantically connected to the real world?
3. *Image Relations*: How are images placed in the visualization?

Reference: The location of the result visualization can be defined in the same way Schmalstieg and Höllerer described the reference of virtual AR objects [SH16]. Based on their work, three kinds of reference are defined in this work:

- *Free Space*: The result visualization is located in the air.
- *Surfaces*: The result is visualized on a physically existing plane.
- *Body*: The result visualization is connected to the user's body.

Result Space: Different possibilities exist for connecting the results semantically to the real world. Three main categories can be differed:

- *Shape:* There is no semantic connection to the real world, images are simply placed on virtual geometrical objects.
- *Metaphor:* The visualization is figuratively embedded into the physical world using real world concepts.
- *Coordinate System:* The visualization connects real world dimensions to image properties.

Image Relations: Datta et al. name different types of image visualizations that can also be used in a composite way (see section 2.1.3) [Dat+08]. These categories are slightly adapted and used to describe different image relations. Instead of *time-ordered*, the more general term *property-ordered* is used in order to cover a broader range of visualizations. The term *clustered* is not used as a characteristic because images can be clustered independently from their underlying relation.
Therefore the following values are used to describe image relations:

- *Relevance-Ordered* images are sorted based on a certain measure that related to the user's query.
- *Property-Ordered* images are sorted based on certain characteristics of the images.
- *Hierarchy-Ordered* images are sorted based on underlying categorical structure.

3D Result Interaction

3D result interaction denotes the actions taken by the user after a result has been displayed by the system. Based on Norman's cognitive model (see section 2.1.3), users have a certain goal in mind and then execute a series of tasks in order to accomplish that goal. For 3D result interaction, every task involves a modality and a part of the result that the user is interacting with. Additionally, the interaction can be characterized by how intentional it actually was. Therefore the following categories are employed to describe 3D result interaction:

- *Purpose*: What is the goal of the user interaction?
- *Modality*: How is the sequence of tasks carried out?
- *Target*: With which part of the result is the user interacting?
- *Explicitness*: How intentional is the interaction?

Purpose: As outlined in section 2.2.2, three main 3D interactions are especially relevant in AR. In the present work they are used to describe the purpose of the result interaction as follows:

- *Selection* means the user selects images in order to manipulate or explore them.
- *Manipulation* means the user moves pictures to another place in order to save them or use them as a new query.
- *Exploration* means the user investigates the result in order to judge the relevance of the pictures.

Modality: This property is the same as the interaction modality described earlier under *interaction modality* of *natural query specification*.

Target: When describing *relevance feedback*, Datta et al. mention region-based approaches as well as semantic labels or groups of images [Dat+08]. To describe the target of interaction more generally, a gradation similar to the one proposed for queries is used as follows:

- A *Region-based* interaction targets spatial parts of images.
- An *Image-based* interaction targets individual images and their properties.
- A *Group-based* interaction targets groups of images by position or semantics.

Explicitness: Whenever an interaction between user and system happens, it can be more or less intentionally. When a user gazes at a certain result image, additional information might be provided without the *explicit* intention of the user. Such a kind of interaction would be called *implicit*. The two terms are defined as follows:

- An *implicit* user interaction is unintentional and triggers an unexpected result.
- An *explicit* user interaction is intentional and triggers an expected result.

2.3.3 Summary of the Taxonomy

A taxonomy has been proposed in order to give the field of image retrieval within AR some structure to review related work and elaborate concepts in chapter 3. The first part of the taxonomy is called *session parameters* and is used to describe basic properties that characterize user, data and context. The second part is named *interaction process* and structures the steps of interaction that describe the retrieval process during in its course.

As depicted in figure 2.5, the classification for the *user* includes the *goal*, *strategy* and *expertise* of the user. *Data* has been grouped regarding its *size* and *structure* as well as the *domain breadth* and the *expertise breadth*. *Location* and *timespan* have been chosen as determining features of the *context*.

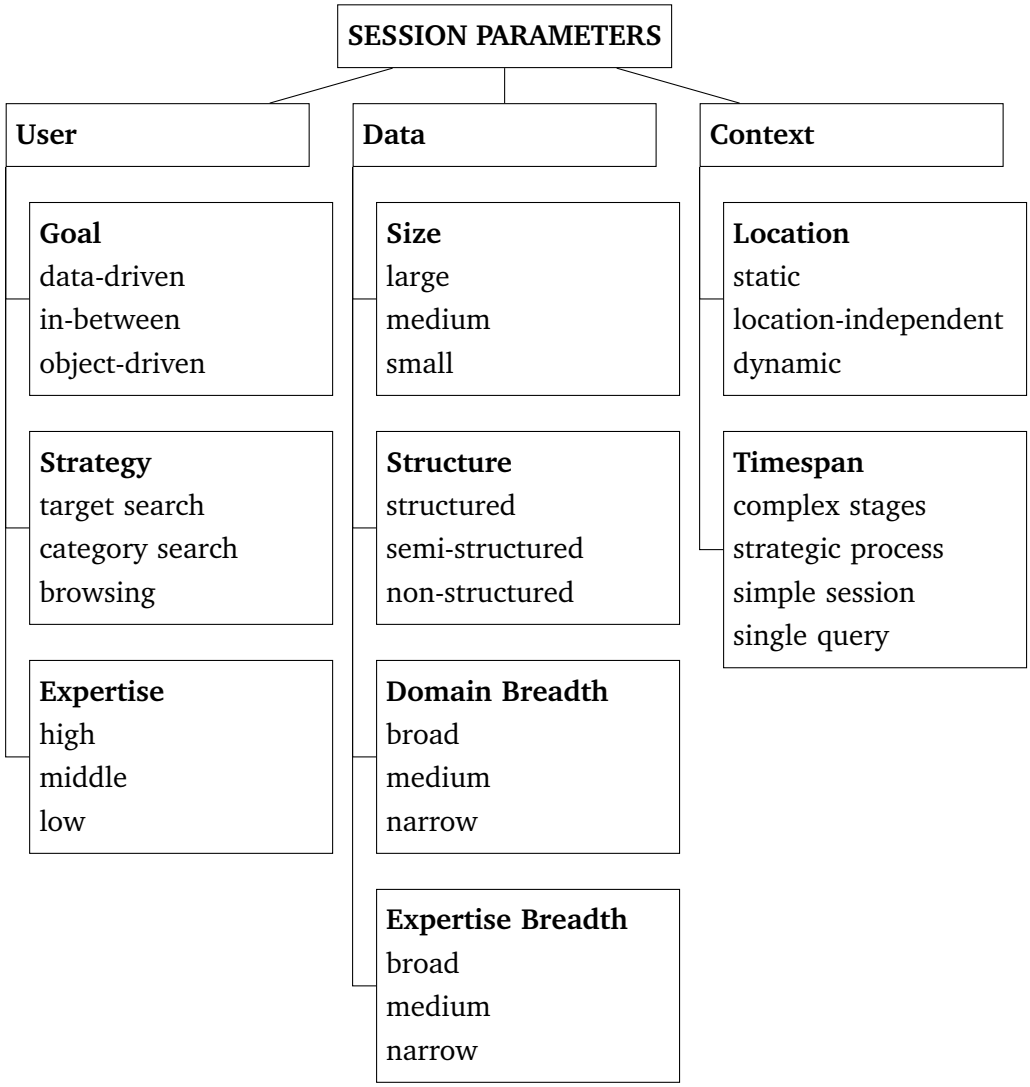


Fig. 2.5: Session parameters part of the proposed taxonomy

The interaction process has been divided into *natural query specification*, *situated result visualization* and *3D result interaction*. For natural query specification, the *target* and the *modality* of the query as well as the *modality of the interaction* and the *effort* have been described. The situated result visualization has been divided according to its *reference*, the semantics of the *result space* and the *relations between the images*. *Modality*, *target*, *purpose* and *explicitness* have been determined as characteristics of 3D result interaction. The full interaction process taxonomy is depicted in figure 2.6.

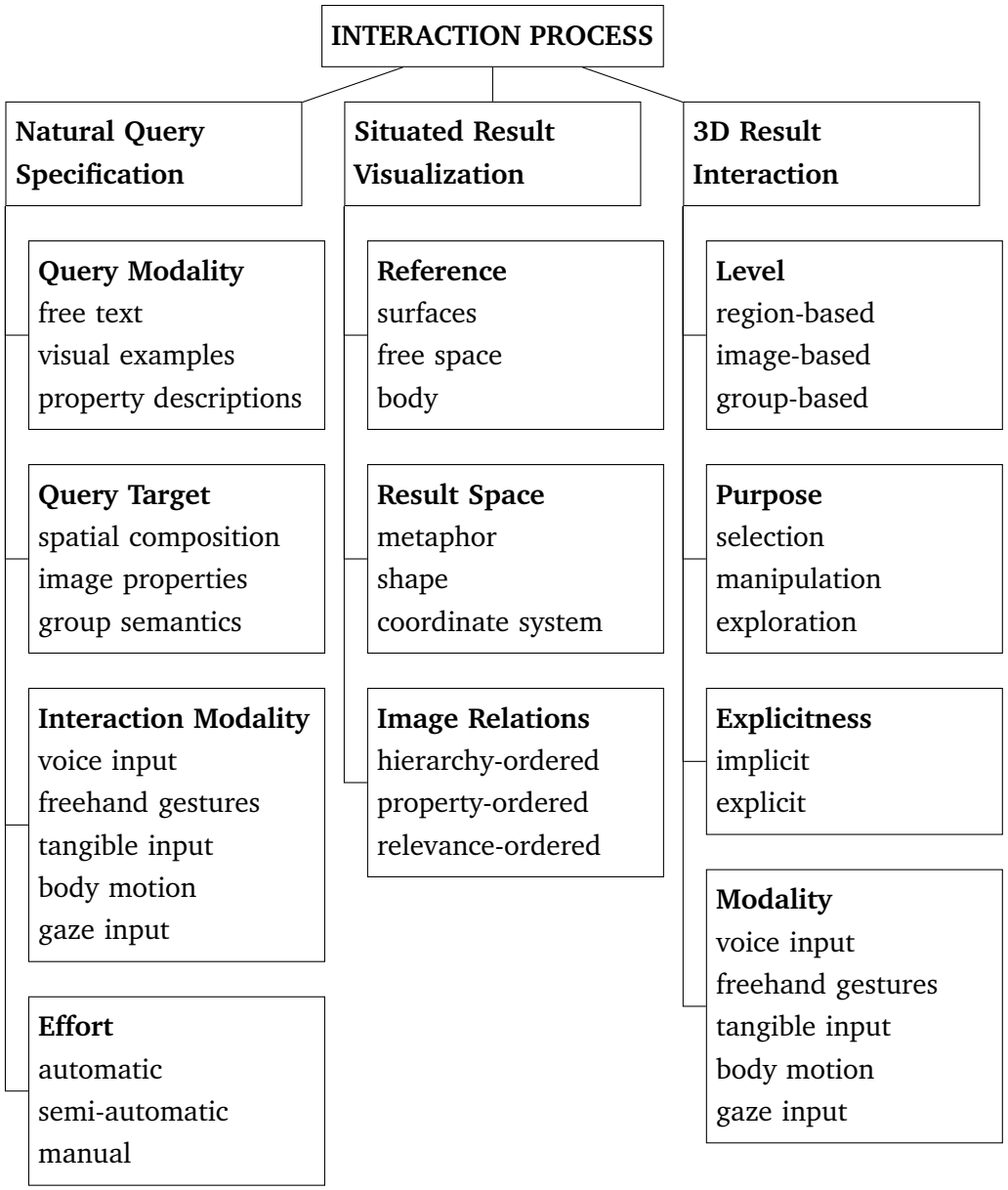


Fig. 2.6: Interaction process part of the taxonomy

3 Concepts for Image Retrieval within Augmented Reality

To evolve elaborated concepts for image retrieval within AR, a series of steps is taken in this chapter. Firstly, chosen related work is reviewed in section 3.1 as a basis for inspiration. A series of rough ideas for interaction is then developed in section 3.2. In section 3.3 requirements for the desired concepts are determined based on the design foundations of chapter 2. Finally in section 3.4, two comprehensive concepts are presented based on the previous sections and evaluated using the requirements.

3.1 Related Work

Due to the nature of combining two big areas of research, it is simply out of scope for this work to review image retrieval and AR in detail. Instead, this section elaborates only on research work with special potential for image retrieval in AR. This includes:

1. Image retrieval concepts that incorporate novel kinds of interaction or visualization.
2. Information retrieval concepts that employ natural interaction.
3. AR browsers that facilitate a kind of information retrieval.

Based on the taxonomy framed in section 2.3, related work is divided into concepts for Natural Query Specification (section 3.1.1), Situated Result Visualization (section 3.1.2) and 3D Result Interaction (section 3.1.3).

3.1.1 Natural Query Specification

Different image retrieval systems are often referred to by their type of query. Therefore in this section the query modality as defined in the taxonomy (see section 2.3) is used to categorize different related work.

Free Text: Natural Language Queries

An advancement of classical keyword search is free text search. This kind of information retrieval paradigm is closely related to natural language queries. Natural

language queries have been a topic in research for a long time, like the work of Harris [Har77] and Kaplan and Jerrold [Kap82] shows. Since then, natural language interfaces have been developed to the extend that they can be used for query relational databases and consumer products with services like Google Assistant, Siri and Cortana [KC13] [LJ14].
Natural language interfaces have also been used for special image retrieval applications. In their 1997 Paper, Harada et al. propose a natural language interface that matches subjective expressions like "show me a cute one" or "Is there a slightly darker one?" with color features [Har+97]. In a study they conducted they found out that their interface does not differ greatly from a graphical color picker in terms of success rate, number of queries and thinking time. However a questionnaire showed that users found the natural language interface more difficult to use because of restrictions in sentence pattern and the usage of the keyboard [Har+97]. Though they state that the use of comparative expressions leads to a high success rate and that the natural language interface is especially suitable for more abstract queries and discovery of unexpected images.

Visual Examples: Photographs and Advanced Sketches

The use of visual examples for querying image retrieval systems has been state-of-the-art in research for a reasonable time now. In 2000, Assfalg and Pala named four different paradigms that have already been state-of-the-art in 2000 [AP00]:

- Example *images* used to retrieve images with global color or texture similarity.
- *Painted* color patches used to retrieve images with certain colors in certain regions.
- *Sketched* object outlines used to retrieve images with certain objects.
- Spatially arranged *icons* used to retrieve images with certain arrangements of concepts.

Tab. 3.1: Drawbacks of visual example paradigms according to Assfalg and Pala [AP00]

Drawback	Image	Painting	Sketching	Iconic
No content editing	x			x
Examples don't fit requirements	x			
Too many examples required	x			x
Drawing abilities required		x	x	
No visual feedback				x
Visual memory required		x	x	x

As pointed out in table 3.1 by Assfalg and Pala, each of those paradigms has several drawbacks [AP00]. To overcome these drawbacks, two general approaches can be identified: Combining multiple existing paradigms or proposing novel paradigms.

In order to focus on promising interaction techniques and paradigms, only related work belonging to these categories is being reviewed while standard image, painting, sketching or iconic interfaces are left out.

Advanced paintings and sketches: Different work has been done in research in order to improve and combine existing interaction paradigms. Table 3.2 gives an overview of how several research papers combine different paradigms or use novel interaction modalities.
In their *Epic* system from 1998, Jose et al. allow users to draw rectangles on a sketchpad and label them semantically [Jos+98]. This kind of query procedure was rated significantly better in all aspects compared to a keyword-based retrieval.
In 2002 Ko and Byun proposed query-by-gesture as a new query method [KB02]. Their system *FRIP* allows users to draw shapes into the air with their hands. The system then retrieves images with objects that are similar to these shapes.
Matkovic et al. used a tangible interface with physical blocks in different colors to let the user create a coarse image composition [Mat+04]. In that way they avoid sketches being too detailed for the algorithm as well as users not wanting to use the interface because they are afraid they can't draw. In a study, most of the users found the tangible interface more likable than a conventional one [Mat+04].
Engel et al. enriched their sketch-based interface with semantical brushes, allowing the user to add a meaning to her patches in the sketch [Eng+11].
Sugimura et al. proposed a sketch-based interface that allows users to transfer colors and patches from retrieved example images into their sketch in order to improve search results [Sug+16].

Tab. 3.2: Overview of paradigms used in related sketch-based interfaces

Name	Image	Paint	Sketch	Icon	Natural interaction
Epic [Jos+98]			x	x	
FRIP [KB02]		x	x		x
Tang. Image Query [Mat+04]		x			x
Semantic Sketches [Eng+11]		x		x	
Sugimura et al. [Sug+16]	x	x			

Photographs: In their work, Assfalg et al. propose a novel type of querying called *Query-By-Photograph* [Ass+98]. Their application allows the user to navigate in a virtual 3D environment and take photographs of the scene. These pictures are then used as an example image to query a database. The 3D scene can be alternated by changing objects, colors and textures. This allows content editing without requiring painting or sketching abilities [Ass+98]. It also allows to cover a wide range of expertise levels: I simple snapshot is possible as well as a professional set-up of the scene. In a user study they found out that their interface was more interesting and

allowed more customization of the query, but was harder to learn than query-by-image and query-by-sketch interfaces [Ass+02].
With the rise of modern mobile phones, the photograph paradigm has also been used to take pictures in the real world and use them to query retrieval systems. In 2006, Jia et al. proposed their system *Photo-to-Search* that allows the user to take snapshots of interested objects and retrieve information about them [Jia+06]. While still being a part of research, mobile visual search has also become part of commercial applications like Google Goggles later on [Gir+11].

Property Descriptions

Apart from text and visual examples, another popular method to query databases is to describe the desired properties of the image. This paradigm of exploring feature space can be applied in two different ways: *Explicitly* by user input through panels, buttons and dials and *implicitly* by tracking the user's interactions. The explicit approach can be found in faceted search interfaces, which are either implemented as graphical user interface (GUI) or tangible user interface (TUI). The second way is used in location-based browsing, where queries are specified through the location and heading of the user.

Faceted search: Faceted search makes use of different image properties and metadata that users can describe in order to retrieve desired images. In their VISMap GUI, Chen and Chang provide areas with different functionality, where each area contains a set of tools [CC01]. These include filters, classifiers, templates, browsers and canvases. Tools can be used to perform range queries, adjust range threshold values and create boolean queries [CC01]. Yee et al. proposed a category based faceted image retrieval system for a collection of 35000 images [Yee+03]. Although their application was slower than a standard system, it was strongly preferred by the users because they found it easier to use, more flexible and insightful concerning information about the image collection [Yee+03].
Faceted search has also been used widely with TUIs. Ullmer et al. introduced tangible query interfaces in 2003, using the so called token+constraint approach in order to let user's specify certain parameters [Ull+03]. Another approach in the form of TUIs is the one of constructive assemblies, where tangibles are used in a constructive way in order to build queries. In doing so, different arrangements of tangibles imply different conjunctions of parameters in boolean queries. Jetter et al. proposed Facet Streams where tokens with certain parameters are used to narrow down the stream of results [Jet+11]. Other constructive queries are built by stacks of tangibles [Klu+12] or two-dimensional arrangements of tangibles on surfaces [Lan+14]. Because of their characteristics, TUIs are especially well suited for collaborative construction of queries [Bla+04].

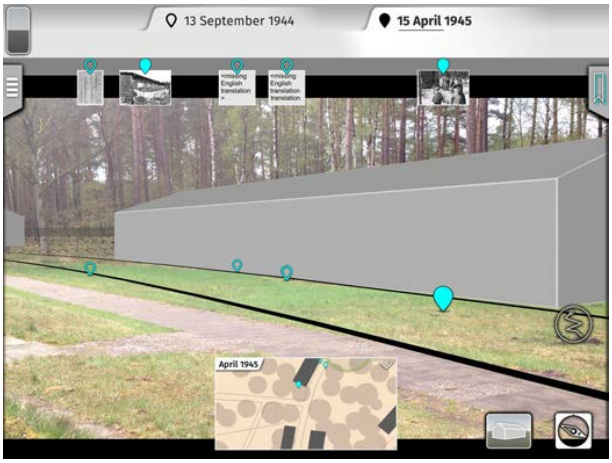


Fig. 3.1: AR field view in the application proposed by Pacheco et al. [Pac+15]

Location-based browsing: In the field of mobile and wearable devices, location-based browsing is used as a paradigm for users to retrieve information.
On their mobile device interface, Jesus et al. allow geographic image retrieval by selecting parts of a map or cardinal directions in order to pose region queries or direction queries respectively [Jes+07]. These can be combined with each other or with semantic concept and example image queries [Jes+07].
Due to its tight coupling with the real world, location-based browsing is also the most prevalent form of information retrieval in mobile AR applications. Commercial applications like Wikitude or Layar are mostly used for general content browsing and navigation, with products, museums, games and advertising being further application areas [Gru+11].
In research, the applications given above are often looked at from an educational point of view (e.g. getting to know historical points of interest). Van Aart et al. employ a location-aware semantic search with minimal interaction in their mobile cultural heritage guide [Aar+10]. Users are able to setup different facets, point to specific directions and select displayed points of interest in order to retrieve historical information in AR. In a similar way, Pacheco et al. provide an application in order to retrieve historical information using two different view modes: Map view and field view [Pac+15]. In map view, either a historical map or a current map can be shown. In field view, the scene is shown either in VR or in AR (see figure 3.1. Content is grouped in points of interest an can be accessed when the user is in a certain radius of the point [Pac+15].
Beneath touristic applications, professional usage of location-based browsing is another field of use in research. This contains for example learning how to operate machines or monitoring states of a system or building. For example the projection-based *iHelmet* proposed by Yeh et al. allows to retrieve detailed visual information about a building using on-site AR [Yeh+12]. Input is realized through positioning of the user and multi-touch gestures on a head-mounted mobile device while output is implemented with the usage of a head-mounted projector.

3.1.2 Situated Result Visualization

Three main categories of result visualizations in AR image retrieval have been identified in section 2.3: Visualizations connecting results to the real world using *metaphors*, *coordinate systems* or a simple *shape*. Related work that contributes to these three categories is reviewed in the following sections.

Metaphorical Visualization

One way of visualizing results in 3D is to use real-life metaphors to give the images a structure that can be embedded into the physical world.
With their *Virgilio* system proposed in 1997, Massari et al. provide a non-immersive VR application that allows to browse multimedia databases metaphorically [Mas+97]. The hierarchical data is thereby transferred into a 3D scene consisting of floors and rooms that can be traversed. Metaphors can be used as following:

- *building:* whole dataset
- *elevator buttons:* set of music types
- *floor:* music type
- *corridor:* set of singers
- *room:* singer
- *door label:* name of the singer
- *photo frame:* image of the singer

In *MediaMetro*, Chiu et al. use a 3D city metaphor for displaying and retrieving multimedia content [Chi+05]. Multimedia documents are represented as ashlar-formed buildings which are grouped in blocks of houses according to the folder they belong to. Different sites of each building show different media types of the corresponding document, for example a keyframe on top, storyboards on the facade and slides with text on the sides [Chi+05].
Zhang et al. use a solar system metaphor for their 3D image browsing system [Zha+14]. Each of the planets in the solar system contains a set of semantically similar images which are shown on the surface of the planet at close range and represented by a single image from the distance. This concept was aimed at making image retrieval more fun and got positive feedback from users [Zha+14].
In their BioAR application, Barreiros et al. are using a tree metaphor to visualize machine states as AR overlay [Bar+16]. In their study, users were able to detect a target tree with different shape or color pre-attentively, making significantly more errors in the latter condition [Bar+16]. A possible explanation of the authors is that the real-world background might interfere with the visualization and make color variations less distinguishable. This explanation also accords with the design problems that have been identified in section 2.2.2.

Coordinate Systems

Another way of 3D image visualization is to augment a coordinate system into the physical world and place images into the system according to different properties. This means that the distance between images is actually meaningful. Different solutions are reviewed in this section: Some employ actual 3D visualization while others use 2D interfaces with promising concepts.

Tian and Taylor use a 3D interface to visualize an image database by texture or color, whereby single images are represented as small textured spheres in space [TT00].
Nakazato and Huang propose a VR interface named *3D Mars* where images are placed into space according to color, texture and structure [NH01]. Displayed images always face the user and can optionally be displayed as spheres in order to facilitate the examination of clusters [NH01].
The *MIAOW* image browser by Gomi and Itoh categorizes images by their shooting locations and times and places the resulting clusters into a 3D coordinate system projected on a plane [GI10]. Thereby the x- and y-axes indicate the longitude and latitude respectively while the z-axis displays the time. Users can switch interactively between XY-, XZ- and YZ-planes and zoom into clusters of images. In a study their system was much more effective than a standard image browser [GI10].
Snavely et al. present a 3D interface that allows users to browse images by their shooting location and direction [Sna+06]. Images are placed around a scene and can be explored in free flight, sequentially based on spatial relations or based on objects on photographs. Virtual camera motions and view interpolation is used for smooth transitions [Sna+06].

The VISMap application proposed by Chen and Chang offers different two-dimensional views of retrieved images [CC01]: On *distance map*, results are plotted in a two-dimensional map according to two user-chosen query values while user-chosen thresholds are displayed as rectangles. On *concept map*, query components are displayed as labeled rectangles at a user-chosen position and results are plotted as squares in between them, with the distance to a query component signifying the similarity to the latter (see figure 3.2).

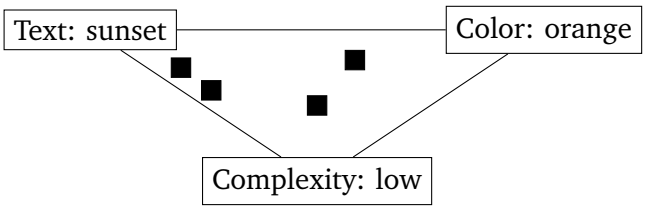


Fig. 3.2: Concept map in VISMap application by Chen and Chang [CC01]

Shapes: Rings, Spheres and Cylinders

Using this kind of visualization, images are placed onto the surface of three-dimensional objects to facilitate perception and browsing of the results. In contrast to metaphors and coordinate systems, mere shapes don't give objects or space in the physical world a meaning, which trades in semantic dimensions for more flexibility. Different research work proposes algorithms that place images into arbitrary shapes like 2D grids or 3D spheres based on different similarity measures [Qua+10] [Fri+15]. Other shapes that have been part of research are for example rings and cylinders.

Rings and Spheres: In user studies for know-item search, Schoeffmann et al. compared 3D rings and globes with 2D grids on different platforms (PC, tablet and smartphone) [Sch+14]. The 3D interfaces were significantly faster and more fun to use on larger screens, but the grid performed better on the smartphone.
Gerald Schaefer proposes a browsing environment where images are placed onto a 3D sphere based on their hue and value properties [Sch10].
In order to browse large image collections on touch-enabled devices, Klaus Schoeffmann proposes a *stack-of-rings* visualization that allows the user to scroll individual rings as well as whole stacks [Sch14]. When searching for a desired image, his interface performed significantly faster than a commonly used grid interface and was preferred by a majority of study participants [Sch14].
Kozma et al. propose an interface where pictures are arranged in vertical rings that can be traversed like a tunnel [Koz+09].

Cylinders: Schoefmann et al. compared a lying 3D cylinder visualization with classical 2D grid-based ones and found out that the 3D visualization had the best workload ratings in five of six categories, with 15 from 28 study participants ranking the 3D interface first overall [Sch+13].
Christmann et al. compared the visualization of images on the outside and inside of a three-dimensional cylinder when looking for unknown photos meeting certain criteria and known photos with familiar locations [Chr+10]. The inner view was more effective and efficient in the first task and was preferred overall by more users. All but one participant of the study rated the 3D interfaces positively in comparison to a standard 2D view [Chr+10].
Van Nieuwenhuizen et al. compared cylindrical 3D interfaces on HMDs and came to the conclusion that placing the user inside the cylinder maximizes user experience [Nie+14]. Additionally, they compared different exploring mechanisms and found out that combining head-tracking with scrolling leads to the best user experience in that regard. In performance measures, no significant differences were traced in the different settings [Nie+14].

3.1.3 3D Result Interaction

Interacting with retrieved results has been a topic in research especially in context of relevance feedback (see section 2.1.3). According to the taxonomy elaborated in section 2.3, three different levels of the result can be targeted with a user interaction: A *region* in a particular image, a particular *image* with its properties or content, or a *group* of images. To facilitate these interactions, different modalities are used in literature. They can be coarsely grouped into the following categories:

- Marking images or parts of them
- Giving linguistic feedback
- Exploring the result with movement

Table 3.3 gives an overview of these modalities by comparing related classic and novel interactions and giving examples for region-, image- and group-based systems in literature, which are explained more detailed in the following sections.

Tab. 3.3: Overview of research work related to 3D result interaction

	Marking	**Language**	**Movement**
Explicitness	explicit	explicit	explicit or implicit
Classic Interaction	mouse gestures	keyboard input	changing virtual view
Novel Interaction	freehand gestures	voice input	gazing, body movement
Region-Based Related Work	Photo Explorer [Sna+06], Interactive Image Retrieval [Jia+15]	Natural Language Object Retrieval [Hu+16]	Faro et al. [Far+10], Papadopoulos et al. [Pap+14]
Image-Based Related Work	3D Mars [NH01]	WhittleSearch [Kov+15], Harada et al. [Har+97]	GaZIR [Koz+09], Keil et al. [Kei+13]
Group-Based Related Work	ImageGrouper [Nak+03], MediaFaces [Zwo+10]		Photo Explorer [Sna+06]

Marking Regions, Images and Groups

Marking images as relevant or irrelevant is a commonly used technique for relevance feedback. However as the present work aims to elaborate novel natural ways of interaction, this section only covers related work that exhibits certain potential for such interaction.

In *ImageGrouper* by Nakazato et al., users interact with the images by grouping them [Nak+03]. Simple operations like drag-and-drop and drawing rectangles allows

fluid and fast interaction in order to query, annotate or organize groups in the form of icons [Nak+03].
In *3D Mars*, images displayed in 3D result space can be marked as relevant or irrelevant in order to improve the query [NH01].
The *photo explorer* proposed by Snavely et al. allows users to mark an object of interest in one picture, which triggers a change to the photograph that provides the best view on the marked object [Sna+06].
In *MediaFaces*, van Zwol et al. provide a system that generates facets for textual image search queries by analyzing search query logs and Flickr tagging behavior. Users can select a facet in order to refine their query [Zwo+10].
In their *interactive image retrieval* system, Jian et al. offer the possibility to mark arbitrary regions in images as foreground and background in order to tell the system in which regions of the image are of particular interest [Jia+15].

Linguistic Feedback on Regions and Image Properties

Language has been a popular way of giving feedback for a long time now. In 1997, Harada et al. presented an *natural language interface* that allows users to express their desired image properties in full sentences using subjective expressions like "cute", "simple" or "warm" [Har+97]. When images have been retrieved, the result can be refined by comparable statements.
With *WhittleSearch* Kovashka et al. provide an interface that lets users refine queries based on properties of the image content by stating comparable sentences [Kov+15]. This is done by either letting the user freely express her wishes based on currently retrieved images (for example: "sportier shoes than on image 3") or by engaging her into a "game" of 20 questions that the system chooses based on its current knowledge. The former approach depends heavily on the quality of the user-chosen feedback while the latter might prove useful if the user wants to spend less time in the process [Kov+15].
In 2016, Hu et al. proposed a system that uses natural language to localize a target object within a given image [Hu+16]. Their study has shown that incorporating spatial configuration and global context leads to a significantly increased performance of *natural language object retrieval*.

Gazing and Moving the View

Movement is being used in result interaction in two different ways: Explicitly by moving the view of the scene or implicitly by gazing at the scene.
The *photo explorer* of Snavely et al. allows users to browse pictures by moving a virtual camera through a 3D scene [Sna+06].

The *spatial interaction techniques* proposed by Keil et al. for AR include seamless information layers around an object that can be explored by physical movement of the user [Kei+13]. Thereby the distance to the object determines the level of detail that is shown, while the angle dictates which kind of information is displayed [Kei+13].
The *GaZIR* interface combines gaze input with a sequence of rings of images [Koz+09]. According to the authors, a grid would induce the users to look at the images in a row-by-row manner, reducing the usefulness of the captured gaze data. Users can zoom through the rings which causes new images to appear based on the gaze data collected on the previous rings [Koz+09].
In [Far+10], users are conducting a simple keyword search that displays a 2D grid view as a result. While observing the result, the gaze of the user is captured and used to re-rank the retrieved images based on similarity. In a user study, 86% of the participants were more satisfied by the re-ranking in comparison to the initial result [Far+10].
Papadopoulos et al. combine a region-based gaze relevance feedback approach with an object-based relevance feedback mechanism [Pap+14]. Their user interface displays a set of ten images that can be replaced with the next set through keyboard interaction. Whenever a user looks at an image for a certain time, the image is displayed bigger in the middle of the interface. In zoomed mode, regions of the image that the user looks at are collected and used to improve the retrieved results [Pap+14]. Another recent example for re-ranking images based on eye tracking can be found in [Li+16].

3.1.4 Summary of Related Work

In this section an overview of related work in information retrieval with focus on image retrieval and AR has been given. The presented work has been classified into the three main steps of image retrieval interaction as defined by the taxonomy in section 2.3. For natural query specification, five different paradigms have been identified: natural language queries, query-by-photograph, advanced sketching, faceted browsing and location-based browsing. For situated result visualization, different simple shapes have been contrasted with coordinate systems and metaphorical visualizations. For 3D result interaction, interactions on regions, images and groups of images by the means of language, marking and movement have been examined. Related work acts as a source of inspiration to elaborate different basic concepts for image retrieval in section 3.2.

3.2 Basic Interaction Concepts for Image Retrieval in Augmented Reality

As a preparation for more comprehensive concepts, a brainstorming has been conducted in order to gather basic interaction concepts that can be used for image retrieval within AR. During brainstorming, all three steps of the interaction process taxonomy elaborated in section 2.3 have been taken into account to ensure that a wide variety of basic concepts was worked out. Additionally, different subcategories were chosen for each step in order to provide further variety.
These categories and subcategories also serve as structure for the present section. Each of the basic concepts is explained by a short description, several properties based on the taxonomy elaborated in section 2.3 and related work examined in section 3.1. Additionally, a sketch is provided for each concept where virtual objects are colored and real world is drawn in black and white.

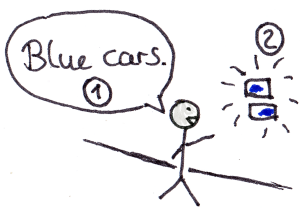
3.2.1 Natural Query Specification

The specification of queries is the first and most important interaction step for any information retrieval session. A number of concepts have been conceived that provide the user with different possibilities to express a query using *free text*, *visual examples* or *property descriptions* as proposed in the taxonomy elaborated in section 2.3.

Free Text

As seen in related work and based on the trend of natural search interfaces, the most widespread use of free text search comes in the form of natural language expressions [Hea11]. For AR, voice input provides a modality that is intuitive and can be combined with freehand or gaze interactions. Table 3.4 depicts one possible usage of free text search in AR.

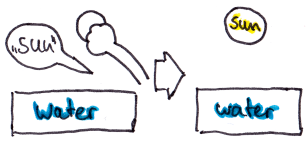
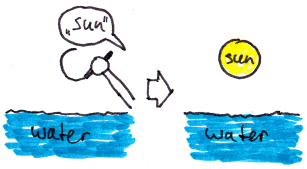
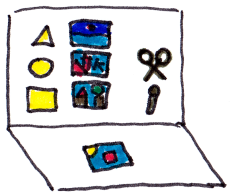
Tab. 3.4: A basic idea for natural query specification using free text

	**Everywhere Query by Voice:** The user specifies a query by telling the system what she searches for. The result appears where the user is pointing or gazing. **Interaction Modality:** Voice Input, Gaze Input, Freehand Gestures **Query Target:** Spatial Composition, Image Properties, Group Semantics **Related Work:** [Har+97] [KC13] [LJ14]

Visual Examples

The most versatile form of querying in image retrieval can be achieved by using visual examples. They are especially useful to describe spatial compositions or image properties. Based on related work, combinations of iconic querying with sketching and example images have been elaborated as concepts and are presented in table 3.5.

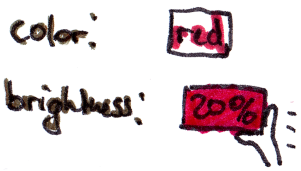
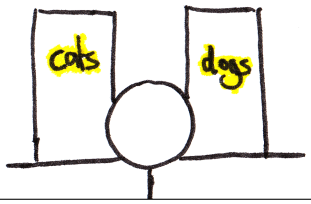
Tab. 3.5: Basic ideas for natural query specification using visual examples

	**Tangible Icon Queries:** The user assigns semantic meaning to tangibles by speaking and constructs the spatial arrangement of the desired image. **Interaction Modality:** Tangible Input, Voice Input **Query Target:** Spatial Composition **Related Work:** [Mat+04] [Eng+11]
	**Freehand Semantic Sketches:** The user draws sketches of the desired image and assigns semantic meaning to parts of it by speaking. **Interaction Modality:** Freehand Gestures, Tangible Input, Voice Input **Query Target:** Spatial Composition **Related Work:** [Jos+98] [KB02]
	**Query by Augmented Photograph:** The user enriches the real world with virtual objects and uses photographs of the scene as a query image. **Interaction Modality:** Freehand Gestures **Query Target:** Spatial Composition, Image Properties **Related Work:** [Ass+02] [Gir+11] [Jia+06]
	**Image Workbench:** A virtual workbench allows the user to construct a query image with the help of shapes, tools and example images. **Interaction Modality:** Freehand Gestures **Query Target:** Spatial Composition, Image Properties **Related Work:** [Eng+11][Sug+16]

Property Descriptions

The third major form of querying as identified in section 2.3 is the description of desired properties. Various related work uses tangible interaction for property descriptions (see section 3.1). In image retrieval, property descriptions are especially useful to describe image properties and group semantics. Table 3.6 shows several concepts that combine tangible interaction with AR for image retrieval.

Tab. 3.6: Basic ideas for natural query specification using property descriptions

	**Tangible Facet Search:** The user arranges tangibles that are assigned with facets in order to construct boolean and range queries. **Interaction Modality:** Tangible Input **Query Target:** Image Properties, Group Semantics **Related Work:** [Yee+03] [Klu+12] [Lan+14]
	**Tangible Image Property:** The user alters property values used for a query by changing the location of the corresponding tangible. **Interaction Modality:** Tangible Input **Query Target:** Image Properties **Related Work:** [CC01] [Ull+03]
	**Real World Hierarchical Browsing:** Users explore an image database by entering different rooms that represent images of different categories **Interaction Modality:** Body Movement **Query Target:** Image Properties, Group Semantics **Related Work:** [Mas+97]

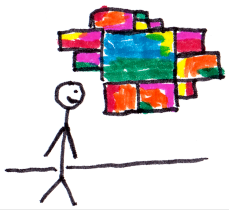
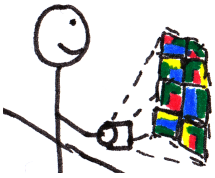
3.2.2 Situated Result Visualization

The visualization of the result is naturally a very important step of image retrieval because it provides the user with the visual information that she is seeking. As section 2.2 points out, AR provides a number of possibilities for visualization but also involves some challenges. The following sections present some concepts that try to make use of AR for image retrieval result visualization taking the named challenges into account. The three main forms of embedding the result space in the real world as identified in section 2.3 are used to structure the present section.

Metaphors

The metaphorical visualization of results is based on real-world paradigms and therefore provides an intuitive connection between the virtual and real worlds. The use of real-world metaphors however might only be applicable in certain settings and hence limit the versatility of the application. Table 3.7 introduces some metaphorical concepts for image retrieval result visualization in AR.

Tab. 3.7: Basic ideas for situated result visualization using metaphors

	**Furniture Based Clustering:** Images are virtually arranged into furniture based on different properties. **Reference:** Surfaces **Image Relations:** Hierarchy-Ordered, Property-Ordered **Related Work:** [Bar+16] [Mas+97]
	**Image Gallery Metaphor:** The images are displayed on a wall like in an image gallery. **Reference:** Surfaces **Image Relations:** Hierarchy-Ordered, Relevance-Ordered, Property-Ordered **Related Work:** [Chi+05]
	**Tangible Projector Metaphor:** With the help of a tangible, the user is able to decide where the result is being visualized. **Reference:** Surfaces **Image Relations:** Hierarchical-Ordered, Relevance-Ordered, Property-Ordered **Related Work:** -

Shapes

In contrast to metaphorical visualization, simple shapes provide a weaker connection with the real world but offer more versatility. One possible concept that makes use of simple shapes for image retrieval result visualization is presented in table 3.8.

Tab. 3.8: A basic idea for situated result visualization using shapes

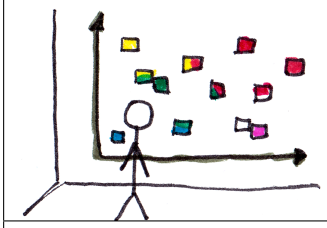
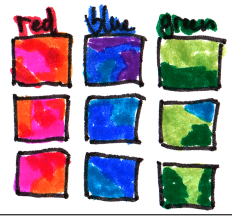
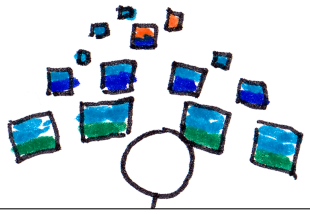
	**Physically Accessible Image Clusters:** Images are collected into clusters in space that the user can explore by physically moving into them. **Reference:** Free Space, Body **Image Relations:** Property-Ordered, Relevance-Ordered **Related Work:** [Sch10] [Sch14] [Nie+14]

Coordinate Systems

The third form of result spaces are coordinate systems. Coordinate systems do need a certain amount of space to be presented on but can provide the user with lots of information about the result and its individual images. These informations can either embody different properties or facets of the visualized images directly or relate to

the user's query by utilizing a similarity measure. Table 3.9 lists different concepts that employ coordinate systems to visualize result images in AR.

Tab. 3.9: Basic ideas for situated result visualization using coordinate systems

	**Real-World Coordinate Systems:** Images are displayed in coordinate systems on real world surfaces or in free space according to different properties. **Reference:** Surfaces, Free Space **Image Relations:** Property-Ordered **Related Work:** [TT00] [CC01] [NH01] [GI10]
	**Facet Matrix:** Images are arranged in matrices according to different facets. **Reference:** Surfaces **Image Relations:** Property-Ordered, Relevance-Ordered **Related Work:** [CC01] [Kei+13]
	**User-Centralized Relevance Visualization:** Images are visualized around the user's position. The nearer images are visualized, the more relevant they are. **Reference:** Body **Image Realtions:** Relevance-Ordered **Related Work:** -

3.2.3 3D Result Interaction

The interaction with the retrieved result is an important last step of image retrieval. As pointed out in section 2.1.3, a number of advanced models of the information seeking process assume that the information need is satisfied in iterative ways and evolves some management of the result. The following sections depict different concepts for region-based, image-based and group-based interaction with the result.

Region-Based Interaction

When retrieving images, users are often interested in certain parts or regions of the image. Therefore the region-based interaction can form an effective way to refine the result in the user's favor. Table 3.10 introduces two ways of natural region-based interaction for image retrieval results.

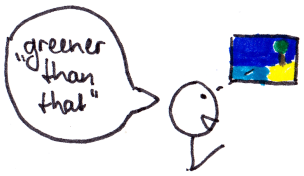
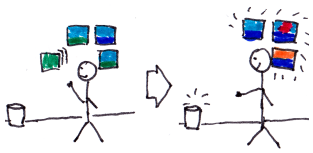
Tab. 3.10: Basic ideas for region-based 3D result interaction

	**Freehand Region Marking:** Users are able to mark relevant regions of images with their hands and give positive or negative feedback using their voice. **Purpose:** Selection, Manipulation **Modality:** Freehand Gestures, Voice Input **Related Work:** [Jia+15]
	**Language Object Feedback:** Users can give feedback about objects in images using their voice, for example "I like that red car on the left". **Purpose:** Selection, Manipulation **Modality:** Voice Input **Related Work:** [Hu+16]

Image-Based Interaction

When users are looking for images with certain properties, the most sufficient way of giving feedback is based on images as a whole. Table 3.11 shows two concepts for image-based result interaction in AR image retrieval.

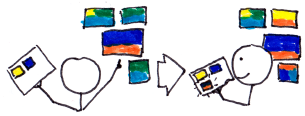
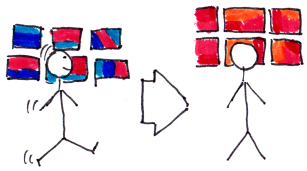
Tab. 3.11: Basic ideas for image-based 3D result interaction

	**Natural Language Image Feedback:** Users can give feedback about images they gaze at in the result in order to express their wishes. **Modality:** Gaze Input, Voice Input **Purpose:** Exploration **Related Work:** [Har+97] [Kov+15] [Hu+16]
	**Metaphorical Relevance Feedback:** Users can give feedback about images by placing them onto real-world objects that symbolize a certain usage. **Modality:** Freehand Gestures **Purpose:** Manipulation **Related Work:** [Mas+97]

Group-Based Interaction

If users are looking for images of a certain category, giving feedback about whole groups of images is a convenient way to refine the result. Table 3.12 presents two concepts for group-based interaction in image retrieval results.

Tab. 3.12: Basic ideas for group-based 3D result interaction

	**Tangible Image Collections:** Users can collect result images on tangible objects in order to give feedback about which pictures they are looking for. **Modality:** Freehand Gestures, Tangible Input **Purpose:** Selection, Manipulation **Related Work:** [Nak+03]
	**Result Refinement by Physical Movement:** While users move along the result, images are changing according to certain properties mapped into space. **Modality:** Body Movement **Purpose:** Exploration **Related Work:** [Kei+13]

3.3 Requirements for Comprehensive Concepts

In order to elaborate and evaluate more comprehensive concepts for image retrieval within AR, different design goals are formulated based on the foundations worked out in chapter 2 and then summarized and embodied in application scenarios. The design goals and scenarios act as a foundation for the elaboration and evaluation of comprehensive concepts in section 3.4.

3.3.1 Design Goals

In order to determine a direction for the composition of comprehensive concepts, different design goals are formulated based on the foundations explored in chapter 2. Taking challenges and design guidelines of information retrieval, image retrieval and AR from sections 2.1.3 and 2.2.2 into account, 13 design goals have been formulated that can be divided into the three main categories *novelty*, *variety* and *usability*. Table 3.13 gives an overview of all design goals and related parts of the taxonomy.

Novelty

In order to depict meaningful research work, the concept should involve some novel aspects of information retrieval and image retrieval as well as AR. More precisely, using location-dynamic AR, the trend of natural search interfaces should be exploited in order to improve image retrieval. The design goals are named as follows:

- *Follow the trend of natural search interfaces* as envisioned by Marti Hearst [Hea11].
- *Provide novel interaction to narrow down the semantic gap* of image retrieval.
- *Establish location-dynamic usage of AR* as defined by the taxonomy under context location.

Tab. 3.13: Summary of design goals

Category	*Domain*	*Taxonomy*
1. Follow the trend of natural search interfaces.		
novelty	information retrieval	-
2. Provide novel interaction to narrow down the semantic gap.		
novelty	image retrieval	-
3. Establish location-dynamic usage of AR.		
novelty	augmented reality	context: location
4. Support different image retrieval strategies.		
variety	image retrieval	user: strategy
5. Support object-driven and data-driven user goals.		
variety	image retrieval	user: goal
6. Support short and complex information retrieval sessions.		
variety	information retrieval	context: timespan
7. Provide means of interaction for practiced and novice users alike.		
variety	information retrieval	user: expertise
8. Allow broad image domains and potential usage outside image retrieval.		
variety	image retrieval	data: domain breadth
9. Avoid data-overload.		
usability	augmented reality	data: size
10. Provide data with temporal coherence.		
usability	augmented reality	context: timespan
11. Show the relation between query and result.		
usability	information retrieval	data: structure
12. Provide query suggestions.		
usability	information retrieval	data: size
13. Support history mechanisms.		
usability	information retrieval	context: timespan

Variety

Because the present work is only a starting point for the conjunction of AR and image retrieval, the elaborated concept should be coarse enough to allow further development into different directions. It should rather be a concept for general interaction rather than for limited application domains. Therefore the design goals are named as follows:

- *Support different image retrieval strategies* as defined by the taxonomy under user strategy.
- *Support object-driven and data-driven user goals* as defined by the taxonomy under user goal.

- *Support short and complex information retrieval sessions* as defined by the taxonomy under context timespan.
- *Provide means of interaction for practiced and novice users alike* to support different levels of expertise as defined by the taxonomy under user expertise.
- *Allow broad image domains and potential usage outside image retrieval* as defined by the taxonomy under data domain breadth.

Usability

In order to form a meaningful concept for interaction and visualization, usability has to be a main goal as well. The concept should therefore follow design guidelines of information retrieval (see section 2.1.3) and overcome challenged of AR. To achieve this goal, different subgoals have been formulated:

- *Avoid data-overload* when displaying information in AR.
- *Provide data with temporal coherence* when displaying information in AR.
- *Show the relation between query and result* to convey data structure.
- *Provide query suggestions* by exploiting data structure.
- *Support history mechanisms* during the retrieval process.

3.3.2 Application Scenarios

For the purpose of complementing and embodying the design goals, four different application scenarios that should be accomplishable with the proposed concept are conceived. Each of the scenarios is described by a short depiction and parameter values characterized by the session part of the taxonomy (see section 2.3.1). Table 3.14 gives an overview of all scenarios and their parameter values.

Paul searches for inspiration

Paul is a 32 year old product designer and needs inspiration for his new work. Because his product should fit into his company's portfolio, he is searching the image database of the company's products for design features which he can pick up.

User: Paul was recently hired by his company and commissioned to design a new product that complements the company's portfolio. His goal is to search the database for images to analyze the visual appearance of the products on them (*data-driven*). As he has no particular goal images in mind while searching, the strategy he chooses is *browsing*. During the process, a *high* level of expertise is necessary in order to identify the design features he is looking for.

Tab. 3.14: Overview of application scenarios and their parameter values

Taxonomy	Paul	Peter	Sarah	Mary
user goal	data-driven	object-driven	in-between	object-driven
user strategy	browsing	browsing	category search	target search
user expertise	high	low	high	medium
data size	medium	large	small	small
data structure	structured	non-structured	structured	semi-structured
domain breadth	narrow	broad	medium	medium
expertise breadth	broad	narrow	broad	medium
timespan	complex stages	simple session	strategic process	single query
location	dynamic	static	static	location-independent

Data: The images of the company's products are stored in a *medium* sized database. Each image is supplemented with rich metadata and semantics, allowing a *structured* access. The images are taken in studio conditions and therefore are very similar in terms of their visual appearance like exposure or resolution (*narrow* domain). In order to understand and adopt the design features of the pictured products, Paul has to apply a great amount of knowledge he has gained over the last years to become an expert (*broad* expertise).

Context: Paul as been given a long-term task by his company and his information retrieval process consists of several *complex stages*. During the process, he might get inspired by his surroundings and pick up ideas wherever he is (*dynamic* location).

Peter wants to decorate his living room

Peter is a 22 year old student who recently moved to a new flat. He wants to search the Internet for a picture that he can order as a poster to decorate his living room.

User: Peter's goal is to find some pictures that fit into his room and suit his taste (*object-driven*). Because he is generally open-minded, he does not search for images of particular categories but rather browses through the web (*browsing*). To identify relevant images, he only needs basic knowledge about his own taste (*low* expertise).

Data: The Internet contains a *large* amount of images that are *non-structured* and very different to each other (*broad* domain). Before searching, Peter thought about

reading up on aesthetics to effortlessly gain expert knowledge, but decided against it (*narrow* expertise).

Context: Peter has found some time to squeeze in a *simple session* of image search while being at home (*static* location).

Sarah picks a picture for an exhibition

Sarah is a 36 year old curator and works for a museum. Recently a place in an image gallery has become vacant and now has to be filled again. Thus Sarah wants to search the databases of images in the inventory of the museum to find a suitable replacement.

User: Sarah's goal is to find an image that suits the gallery visually and semantically (*in-between*). Therefore she searches images that belong to the same era like the neighboring items (*category search*). To find a suiting picture, she has to apply a *high* amount of expertise.

Data: The amount of pictures in the inventory of the museum is comparatively *small* and well *structured*. All the images are paintings but they belong to different eras (*medium* domain breadth). A lot of knowledge that Sarah has gained over the last years has to be applied in order to find the right image (*broad* expertise).

Context: Sarah has to fulfill her task while being *static* at work. She runs through a *strategic process* of image retrieval in which she checks different images until she finds the picture that fits perfectly.

Mary needs an image for her article

Mary is a 28 year old editorial journalist and has just finished an article about the war in Syria in her home office. She has an important meeting coming up at work and needs to find an appropriate image to illustrate her article while traveling to work by train.

User: Mary hast a certain picture in her mind (*target search*) that she wants to use as a striking image for her article (*object-driven*). Because she did a considerable amount of research for her article, she has gained *medium* expertise to find the right picture.

Data: The amount of photographs that are available to Mary is *small*. Because her photographer did not have enough time, only some of the images hold metadata

(*semi-structured*). All the photographs have been shot in the same place, but they are showing different angles of the scene (*medium* domain breadth). Mary has gained her expertise during the last weeks when researching for her article (*medium* expertise breadth).

Context: Mary is currently on the way to work. She has one particular image in mind that she wants to search for independently of her current location (*location-independent*). To achieve this goal, she wants to issue a *single query* that describes the image.

3.4 Comprehensive Concepts

The basic interaction concepts presented in section 3.2 lay the foundation for further elaboration of concepts. They have been analyzed with regard to the concept requirements presented in section 3.3 in order to develop a smaller set of more comprehensive concepts.
During the course of the analysis, two comprehensive concepts emerged: *Tangible Query Workbench* and *Situated Photograph Queries*. The two of them are elaborated and presented in sections 3.4.1 and 3.4.2 respectively. Section 3.4.3 depicts how they conform to the concept requirements shown in section 3.3.

3.4.1 Tangible Query Workbench

One concept family that has stuck out when analyzing the basic concepts presented in section 3.2 is tangible interaction. Related work in section 3.1.1 has shown that tangible search interfaces are well established in research and provide the user with an intuitive way of formulating queries.
The combination of tangible search interfaces with AR allows the augmentation of generic tangibles with specific visual cues that facilitate recognition. Therefore it is possible to use few generic tangibles and still provide a rich set of recognizable interactions.
The idea of the tangible query workbench is to provide the users with a set of powerful tangibles that let them construct image retrieval queries. Using tangible interaction, the composition of images with shapes, textures, colors and semantics is enabled as well as the construction of boolean and range queries to describe certain properties of the image and group semantics.
The concept can be divided into two main components: The *workbench* that is being operated on and the *tangibles* that are used on the workbench (see figure 3.3).

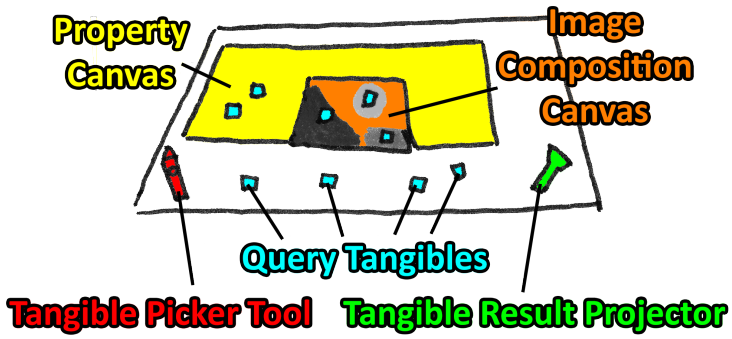


Fig. 3.3: Tangible Query Workbench

Tangibles

There are three types of tangibles that correspond to the three parts of the input-process-output model: A *tangible picker tool* for input, a set of *query tangibles* for processing queries and a *tangible result projector* for output.

Tangible picker tool: The tangible picker tool allows the users to extract properties of the real world into their query. It has the shape of a laser pointer with a square cross-section and a button for the thumb on each side.
While pitch and yaw of the tool determine the direction the user aims at, the roll of the tool can be used to switch between different functions: By rotating the picker tool in their hands, the users are able to switch between picking colors, shapes, textures or semantics. The different sides of the tangible have varying surfaces in order to provide haptic feedback about the current mode of the tool. Additionally the cursor that assists the user with aiming changes its appearance to give visual feedback (see figure 3.4).
The transfer of real world properties onto query tangibles is executed by holding down the laser pointer button, moving the pointer from the real world object to the query tangible, and releasing the button. Rotating the tool before releasing the button enables fine-tuning of the transferred property. This interaction allows to scale shapes, adjust hues or alter semantic numbers like year dates.

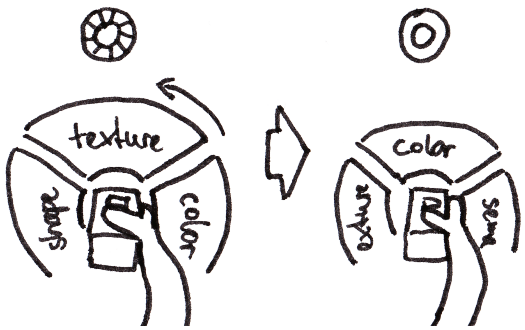


Fig. 3.4: Visual feedback of cursor and menu when operating the tangible picker tool

Query tangibles: Query tangibles have the shape of a coin and represent the properties of the image that the user searches for. They have no meaning in their initial state but can be augmented with colors, shapes, textures and semantics using the tangible picker tool. Table 3.15 shows the possible functions that query tangibles can provide using different tools and canvases.
Whenever an query tangible is placed on either of the zones on the workbench, it contributes to the query. Whenever a query tangible is shaken, it is reset into its initial state and can be newly assigned.

Tab. 3.15: Function of query tangibles with varying tools on different canvases

Tool	On property canvas	On composition canvas
color picker	color filter	painting
shape picker	shape filter	sketching
texture picker	texture filter	patch of example image
semantics picker	keyword query	iconic querying

Tangible result projector: This tangible determines where to display the result that the queries of the user produce. It is shaped and used like a flashlight and employs a projector metaphor to display the result. A button on the tangible allows to activate and deactivate the presentation of the result.

Workbench

As stated by Smeulders et al. and adapted in the taxonomy in section 2.3.2, a query can target spatial composition, image properties and group semantics of an image [Sme+00]. To enable different query targets, the workbench is divided into two zones: On the *image composition canvas*, the user can exploit the advantages of visual query examples by constructing the spatial appearance of the desired image. On the *property canvas*, the user can describe the image properties and group semantics with keywords and tags.

3.4.2 Situated Photograph Queries

Besides tangible interaction the query-by-photograph paradigm by Assfalg et al. is another promising approach to be used in AR [Ass+02]. The basic concept of augmented photographs presented in section 3.2.1 picks up this paradigm and transfers it into AR. It allows the combination of real and virtual objects to create pictures that are used as queries for image retrieval. The comprehensive concept of *situated photograph queries* expands this paradigm by adding semantic search and aspects of physically accessible image clusters and the facet matrix presented

in section 3.2.2. This concepts utilizes two different parts for input and output respectively: A number of *situated queries* and the *result canvas*.

Situated Queries

Situated queries allow the user to extract information from the real world by taking a photograph and then exploring the result using tags and filters. The different steps of interaction during the usage of situated queries are described in the following paragraphs.

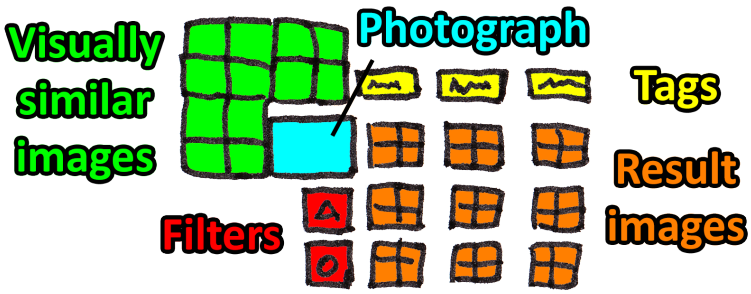


Fig. 3.5: Situated Photograph Queries

Photographing the scene: Situated queries are created by gazing at a scene and performing a shutter release gesture like on a real camera. A zooming gesture can be used to determine the photographed detail of the scene. This is done by moving one hand like when using a zoom ring of a real camera.
The photograph is then displayed in front of the user and augmented with images that are visually similar to it as well as a number of tags that describe it. Visually similar images are displayed to the top left of the photograph while the tags and filters are used to form a matrix of images to the bottom right of the photograph (see figure 3.5).

Exploring the result: The tag matrix consists of columns of different tags that are initially extracted from the user's photograph and rows of different search filters. Search filters can be image properties like colors or metadata. The outcome of this is a matrix of images that are described by the combination of tags and filters, whereby the first row of images is not filtered at all.
Users can explore situated queries by body movement. Figure 3.6 illustrates the different possibilities: When the user moves nearer to the query, the amount of images displayed in the cells of the matrix. When the user gazes to the right, more tags are displayed that describe the image. The opposite happens when the user moves away from the query or gazes to the left. From afar, queries are represented only by their photograph and tags without further images. In this state, queries can hold icons that notify the user of changes inside the result.

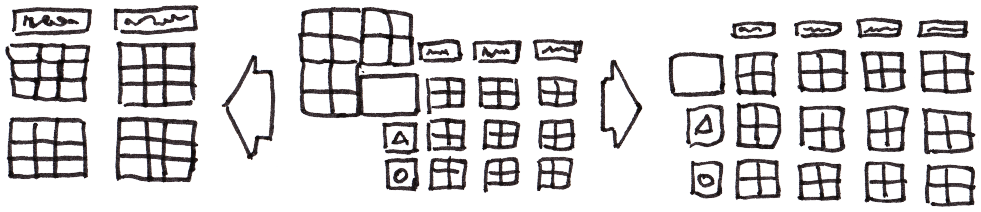


Fig. 3.6: Two ways of exploring the result: Moving nearer to get more result images for certain tags (left) or gazing to the right to get more tags (right)

Refining the result: In order to refine the result, the user can combine different tags or filters into one column or row respectively. This is done by performing a grab gesture while gazing at one tag or filter and then performing a release gesture while gazing at the other tag or filter. A combination of tags or filters can be undone by gazing at the corresponding column or row and performing a slice gesture with one hand.
Tags or filters that are not of interest can be deleted by gazing at them and performing a wipe gesture with one hand. Tags or filters can also be transferred from one query to another. The corresponding interaction is the same like when combining tags or filters, except that the target for the release gesture is now another query.

Result Canvas

Whenever a user chooses an image inside a situated query as a relevant result, it is displayed on the result canvas. The canvas is displayed on a wall near the user and can be connected to any personal device or service in order to save the produced results. In collaboration scenarios, users can have individual situated queries but share one result canvas. The result canvas can be reset by gazing at it and performing a wipe gesture with one hand.

3.4.3 Conformance of Concept Requirements

This section investigates the concepts' conformance to concept requirements. At first, all *application scenarios* are outlined for both concepts, then the concepts are contrasted with regard to the three design goals *novelty*, *variety* and *usability*.

Application Scenarios

In order to approve the proposed usage of both concepts, this section is acting out all of the four application scenarios for either of them.

Tangible Query Workbench: The tangible query workbench can be used in the following ways to achieve the various tasks presented in the application scenarios:
Paul searches for inspiration. He carries his tangible picker tool and several query tangibles with him in order to extract shapes and tags from the real world whenever he sees something inspiring. When being at home or at work he browses through the data by placing query tangibles into workbench zones in different combinations. The result projector allows him to investigate features of the retrieved images on a large wall. Over the time, he refines his query tangibles more and more until he eventually gets the dazzling idea for his new product.
Peter wants to decorate his living room. He uses his tangible picker tool to extract colors from his living room to browse through different images that suit the room and his taste. The tangible result projector allows him to project the image onto its designated place and eventually make a decision.
Sarah picks a picture for an exhibition. She uses her tangible picker tool to extract the the era and the texture of the neighboring pictures onto different query tangibles. When placing the tangibles on the workbench she obtains the category of pictures that she was looking for and projects them onto the designated place in order to evaluate how well they suit. After checking different images of that category she eventually finds a suiting image.
Mary needs an image for her article. She extracts the semantics of her article onto query tangibles but has to wait until she is at work to use the tangible query workbench. There she submits a single query to find the desired image that she was looking for.

Situated Photograph Queries: The situated photograph queries can be used in the following ways to achieve the various tasks presented in the application scenarios:
Paul searches for inspiration. Whenever he discovers an inspiring object, he takes a photograph and chooses visually similar images that are saved on his result canvas at the large wall at work. Over the time, he collects different images in that way and eventually gets the dazzling idea for his new product while browsing through the result canvas.
Peter wants to decorate his living room. He takes a photograph of his living room and extracts the colors that fit his taste. He chooses the images that interest him and views them on the result canvas on his wall in order to make a decision.
Sarah picks a picture for an exhibition. She takes a photograph of the neighboring picture and looks up in the cell of the tag matrix that represents the combination of the picture's era and its characteristic colors. She picks images from the cell and checks on the result canvas how well they work. After checking different images of that category she eventually finds a suiting image.
Mary needs an image for her article. She takes a photograph of her article and combines the extracted tags "Syria" and "war". Looking up in the right year date filter cell, she finds the image that she was looking for and selects it.

Novelty

Because of the usage of AR and HMDs, novelty is a goal that both concepts meet entirely. The tangible query workbench makes use of tangible interaction and narrows down the semantic gap by providing different tools and workspace zones for the user. The concept of situated photograph queries uses gaze and gestures to provide the user with the possibility to extract features from the real world using a photograph metaphor. Table 3.16 gives an overview over the three design goals of the novelty category and how the goals are met by either of the presented concepts.

Tab. 3.16: Design goals: Novelty of the proposed concepts

Goal	**Tangible Query Workbench**	**Situated Photograph Queries**
Follow the trend of natural search interfaces.	tangible interaction in AR	gestures and gazing in AR
Provide novel interaction to narrow down the semantic gap.	rich tangible tools and different workspace zones	real-world tag extraction with photographs
Establish location-dynamic usage of AR.	portable augmented tangibles	physically exploring situated queries

Variety

The variety of the proposed concepts is already embodied into the four different application scenarios that have been played through before. Additionally, table 3.17 illustrates how the different design goals are met, referring to the various application scenarios when necessary.

Usability

The comprehensive concepts were developed with usability as a main target in mind. As table 3.18 points out, they both employ different mechanisms in order to fulfill the set usability goals.

Tab. 3.17: Design goals: Variety of the proposed concepts

Goal	Tangible Query Workbench	Situated Photograph Queries
Support different image retrieval strategies.	target search and category search by extracting desired properties, browsing by altering tangibles	target search and category search by looking up desired cells in the tag matrix, browsing through visually similar images
Support object-driven and data-driven user goals.	result projection can be used to investigate data in pictures and place pictures in the real world like objects	result canvas can be used to investigate data in pictures and place pictures in the real world like objects
Support short and complex information retrieval sessions.	see application scenarios: short sessions (Mary and Peter) are possible as well as long sessions (Sarah and Paul)	
Provide means of interaction for practiced and novice users alike.	tangibles employ well-known metaphors but also offer possibilities for rich combinations	a simple photograph already provides results, combinations of tags and filters offer profound interactions
Allow broad image domains and potential usage outside image retrieval.	extraction of features is not limited to visual aspects	photographing scenes to retrieve information is not limited to images

Tab. 3.18: Design goals: Usability of the proposed concepts

Goal	Tangible Query Workbench	Situated Photograph Queries
Avoid data-overload.	the display of the result can be placed and triggered by the user	the user can determine the level of detail and delete single tags or filters
Provide data with temporal coherence.	result is only produced when tangibles are in the workbench zones	the user can browse the queries with varying level of detail
Show the relation between query and result.	changes of tangibles directly trigger a change of the result	the tag matrix provides an overview over how results emerge from the query
Provide query suggestions.	extracted features embody suggestions	extracted tags and provided filters embody suggestions
Support history mechanisms.	tangibles are used to save features	situated queries are persistent

4 Prototypic Implementation of Situated Photograph Queries

This chapter describes the prototypic implementation that is a part of the present work. The aim of the implementation was to embody chosen concepts into a prototype in order to evaluate them. A Microsoft HoloLens device was chosen for the implementation because it offers a stable tracking and reliable augmentation and provides a number of relevant functions built-in. This allows to concentrate on aspects of interaction and visualization rather than technical aspects of the AR application.
The choice of concept for the implementation had to be made between tangible query workbench and situated photograph queries. An implementation of the tangible query workbench would involve a number of complex operations that are in their combination out of scope for a single thesis. These operations include manufacturing different tangibles, implementing a tracking of tangibles and preparing an image search system that supports compositions of visual aspects and semantics.
The implementation of situated photograph queries however involves no devices other than the HoloLens. Photographing the scene and augmenting it with situated visualizations are functions that can be implemented using built-in functions of the HoloLens device. Therefore situated photograph queries have been chosen as the concept for the implementation.
Section 4.1 describes the process and structure of this implementation, section 4.2 explains the installation and usage of the application and section 4.3 discusses the implementation by comparing it to the original concept and giving an outlook on further possible development.

4.1 Implementation Design

The design of the implementation is explained by first giving an overview of the involved process in section 4.1.1 and then presenting the final structure of the prototype in section 4.1.2.

4.1.1 Implementation Process

The implementation process was conducted as agile iterative development. During the process, the development of chosen features with the HoloLens emulator alternated with the testing of different milestones on the HoloLens device. The

goal of this approach was to iteratively enrich the prototype in order to match the concept more closely with every step without a predefined final implementation state. This procedure was chosen because the HoloLens is a rather new device and it was unclear in the first place how well the implementation process would work and which problems would be encountered.
The following sections outline the functionality of the corresponding prototype and the problems encountered during the development for each of the four milestones.

First Milestone

The prototype of the first milestone contained the core functionality of the application. It allowed the user to take a photograph by executing an air tap gesture and analyzed the created picture with the help of the Microsoft Computer Vision Application programming interface (API). The resulting JavaScript Object Notation (JSON) data was displayed in the console.

Problems during the development: After trying version 5.5.1 of Unity, the usage of the photo mode provoked an error, which is why further development was done entirely in Version 5.4.0.
When using the built-in web request of Unity for sending the photograph to the computer vision API, every request resulted in an "unsupported media" response. Consequently, the *System.Net.Http.HttpClient* class has been used for the request instead. Because Unity does not support the threading that is used by the HttpClient, the code has been escaped from Unity by using preprocessor directives. Additionally, the small utility class *MainThreadExecute* has been implemented and is used to process the result of the HTTP Request back in Unity's main thread.

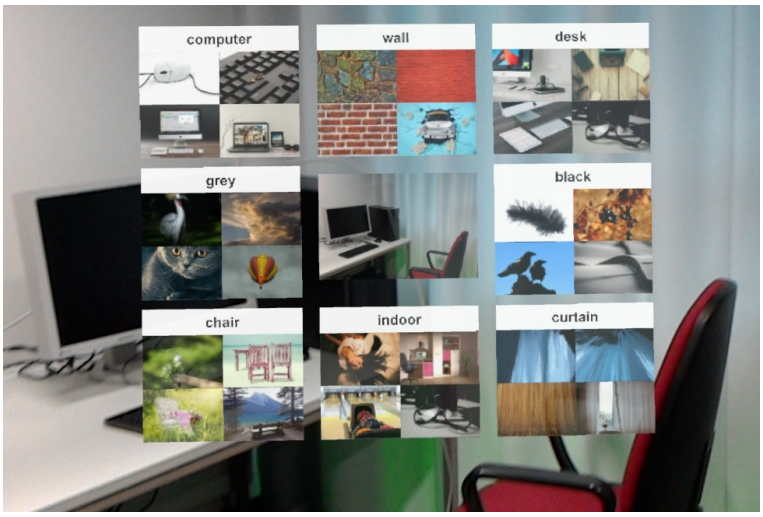


Fig. 4.1: Mixed reality capture of the second prototype

Second Milestone

The prototype of the second milestone included a user interface that displayed tags and corresponding preview images but did not allow further interaction. The tags and preview images automatically disappeared when the distance between user and query exceeded a certain measure. Internally, the final structure of the controllers had been implemented in preparation for further development. Figure 4.1 shows a mixed reality capture of the second prototype.

Problems during the development: The scaling of the Unity UI components has been tricky as it is difficult to estimate the sizes of the elements when using the emulator. During the test of the second prototype, buttons were hard to focus as they were vertically too small. Additionally, the scaling and the placement of the whole interface had to be tweaked as well.

Third Milestone

For the third milestone, interactivity was added to the prototype. The button sizes were adjusted to facilitate air tapping on them. The selection and deselection of tags was implemented together with the canvas that showed the resulting images for the combination of all selected tags (see 4.2). The result canvas was automatically placed on the nearest wall to the user. Resetting the result was enabled by air tapping on the header of the result canvas. Tapping on a preview image enabled to replace the original photograph with the chosen image, which was then being tagged instead. Deleting the query was enabled by air tapping on the query image.

Fig. 4.2: Combination of tags using the third prototype

Problems during the development: A memory problem had been encountered which resulted in a crashing application after a certain amount of browsing and thereby downloading pictures. Although different measures like using smaller image sizes have been taken to tackle this problem, it could not be eliminated completely. Additionally, the automatic placement of the result canvas did not always produce the desired results, sometimes resulting in inappropriate placement of the canvas.

Final Milestone

Because the third prototype depicted a simple but usable implementation of the basic concept of situated photograph queries, the fourth and final milestone was dedicated to refinement and improvement of user experience rather than the implementation of further parts of the original concept.
The automatic placement of the result canvas has been replaced with a manual mode when starting the application. The user can determine the placement of the result canvas by gazing at the desired place and performing an air tap gesture. Whenever the result is reset and has no active tags, the user can activate the result placement again by air tapping on the header of the result canvas.
The automatic minimization of queries has been changed to incorporate the user's direction of gaze. Additionally, the possibility of minimizing queries manually has been provided by air tapping on the query image. The delete function has instead been moved to a button beneath the query image.
An undo function has been added to allow the user to browse backwards when having selected a preview image as new query image. Cursors have been added to provide feedback whenever a photograph can be taken, a button can be tapped or an operation is currently taking place. Figure 4.3 shows a mixed reality capture of the final prototype.

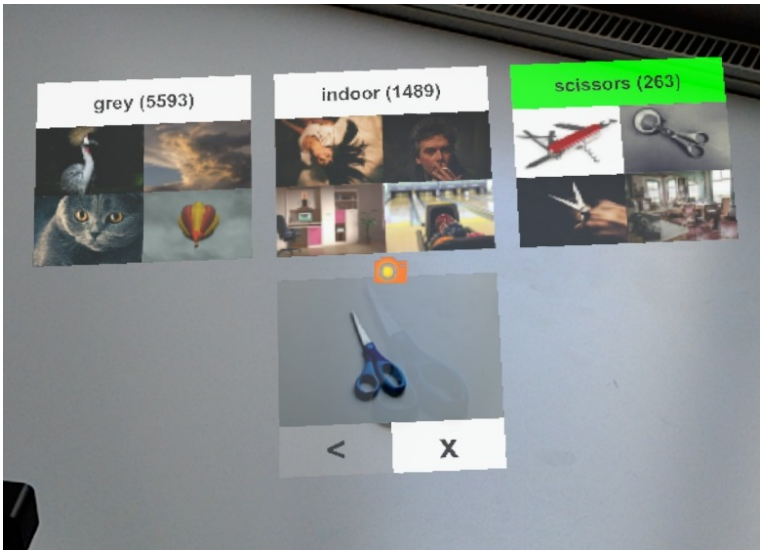


Fig. 4.3: Mixed reality capture of the final prototype

Problems during the development: The usage of the Unity UI components conflicted with the cursors provided in the HoloToolkit. Even after adding a box collider to the UI, the cursor kept jumping back and forth between the front and the back of the UI. Since these kind of problems have been described by other developers on the Internet as well, an own UI cursor has been implemented that is displayed slightly in front of the UI.

Although interaction with the UI components was still possible when parts of the spatial mapping overlapped the interface, the wrong cursor was shown because the raycast collided with the spatial mapping rather than the UI. Therefore a utility class named *ButtonEvents* has been implemented that triggers a boolean to signalize whenever the user gazes at a button.

4.1.2 Structure of the Implementation

The concept of situated photograph queries is meant to be variable and the current prototype depicts only a basic implementation. Therefore the structure of the implementation is kept modular and employs the model view controller (MVC) pattern. This allows to replace existing parts with new ones and to further expand the implementation.
The following sections present external dependencies as well as models, views and controllers that have been created for the present implementation. Figure 4.4 gives an overview of the created classes. A more detailed documentation of classes and methods can be found inside of the code.

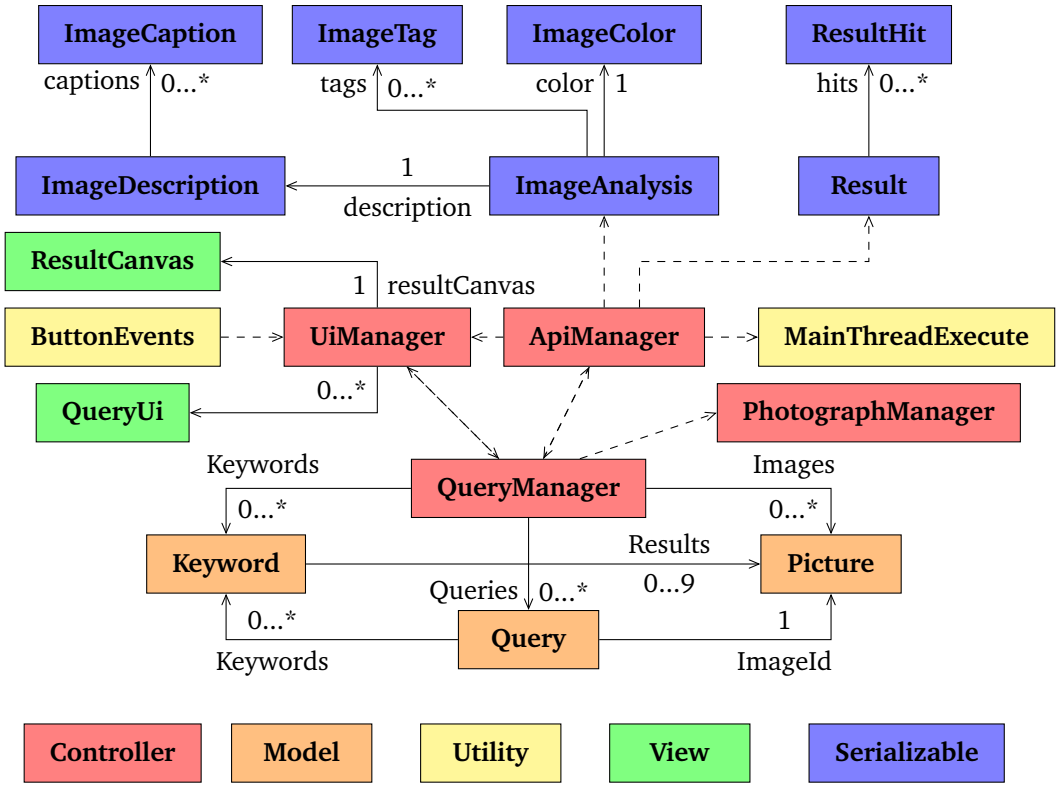


Fig. 4.4: Structure of the final implementation

Application Dependencies

The *Unity Engine* has been chosen as main platform for development because it is the recommended way of getting started with HoloLens development and offers effortless access. Basic system functions and classes of the *.NET Framework* have been used as well. Beneath that Microsoft's *HoloToolkit* supplemented further HoloLens-related functions, namely the *GazeManager* and the *Spatial Mapping Prefab* that have been utilized for the positioning of UI components.
Additionally, the application relies on two different representational state transfer (REST) APIs: Microsoft's *Computer Vision API* has been chosen for the image tagging because it provides reliable results and allows 5.000 free requests per month. The *Pixabay API* has been used for keyword-based image search because it provides effortless access to a big database of free pictures.

Created Views

The views of the created application have been implemented using the UI system of Unity. There are two main views that are used in this application: The QueryUi (see figure 4.5) and the ResultCanvas (see figure 4.6).

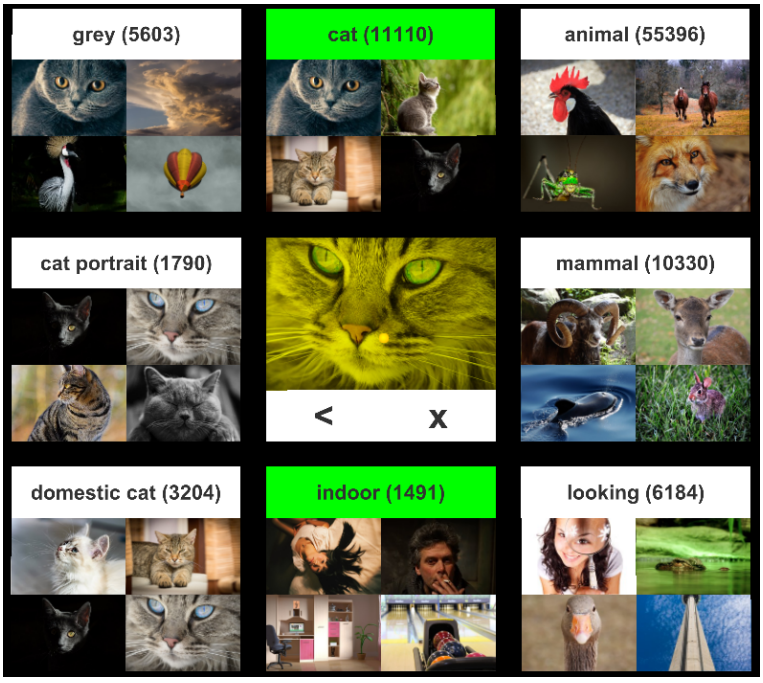


Fig. 4.5: QueryUi

The *QueryUi* is a representation of the situated photograph with tag matrix that has been prepared as a Unity prefab. During runtime, whenever the user takes a photograph, an instance of the QueryUi prefab is created and named after the identifier of the query.

Fig. 4.6: ResultCanvas

The *ResultCanvas* is also created in advance and placed after the start of the application. Both views hold a Unity box collider to interact with the raycast of the Gaze Manager and consist of Unity UI buttons, either with a background image (used to display pictures) or with text (used to display keywords). The buttons are filled with content and activated as and when required. The highlighting of the buttons has been realized using the internal color change function of Unity.
In addition to the main views, three *Cursors* are created using Unity sprites and placed dynamically according to the users gaze and interaction (see figure 4.7).

Fig. 4.7: Cursors used in the application: UiCursor, PhotoCursor and WaintingCursor

Created Models

Three main classes have been created that represent the core forms of data that are used in this application: queries, pictures and keywords.
A *Query* object represents one search request of the user consisting of a query image and a list of keywords. These objects are referenced in the query using a unique identifier. Additionally, a query holds a list of keywords that are currently active (i.e. contributing to the result on the result canvas) and, if the query was created by tapping on a preview image, an identifier of another query that the current query emerged of.
A *Picture* object represents an image that can be displayed as a preview image, result image or query image. Each picture contains a unique identifier, its content in the form of a Unity texture and a list of colors and tags. Additionally, the raw byte data

can be retrieved and a flag indicates if the picture has already been tagged before.
A *Keyword* object represents a tag that can be used as a keyword for searching. It contains the number of results it produces and a list of picture identifiers that represent the result images.
Besides these main classes, the *ImageTag* and *Result* classes have been implemented to provide Unity's JSON serialization a structure for understanding the REST API responses by the Computer Vision API and the Pixabay API respectively.

Created Controllers

The core part of the application are four classes that are implemented as singletons and act as controllers. Figure 4.9 depicts the functional interaction between them when a new query is being created.

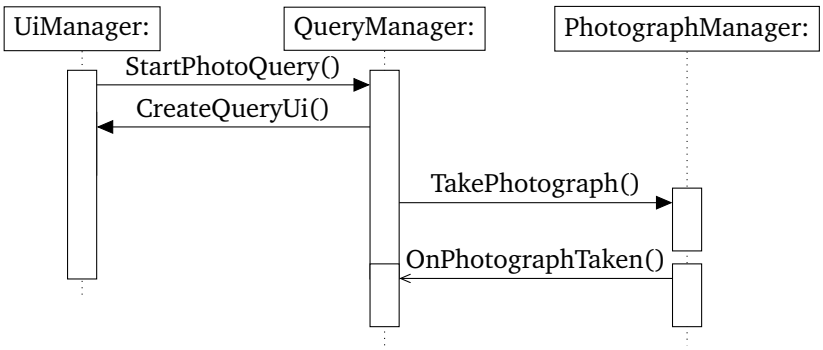


Fig. 4.8: Sequence of taking a photograph (parameters are omitted for reasons of clarity)

The *ApiManager* is responsible for contacting both APIs and preparing the received results for further usage. This includes retrieving tags for images, searching for images and downloading images. The ApiManager fires events when images have been tagged or downloaded.
The *PhotographManager* is responsible for taking photographs and cropping them and fires an event when a photograph has been taken (see figure 4.8).
The *QueryManager* is the core of the application. It holds dictionaries of queries, pictures and keywords and provides methods for retrieving and altering these models.
The *UiManager* detects user input and manages the views described before. This includes placing and viewing cursors as well as filling QueryUis and the ResultCanvas with content. Additionally, a simple logging function can be used during studies to monitor user actions in the HoloLens device portal.

Besides these controllers, two utility classes have been implemented: *MainThreadExecute* is used to process the result of the Computer Vision API request in Unity's main thread and *ButtonEvents* signalizes when users enters or leaves a button with their gaze (see section 4.1.1 for a more detailed elaboration on these workarounds).

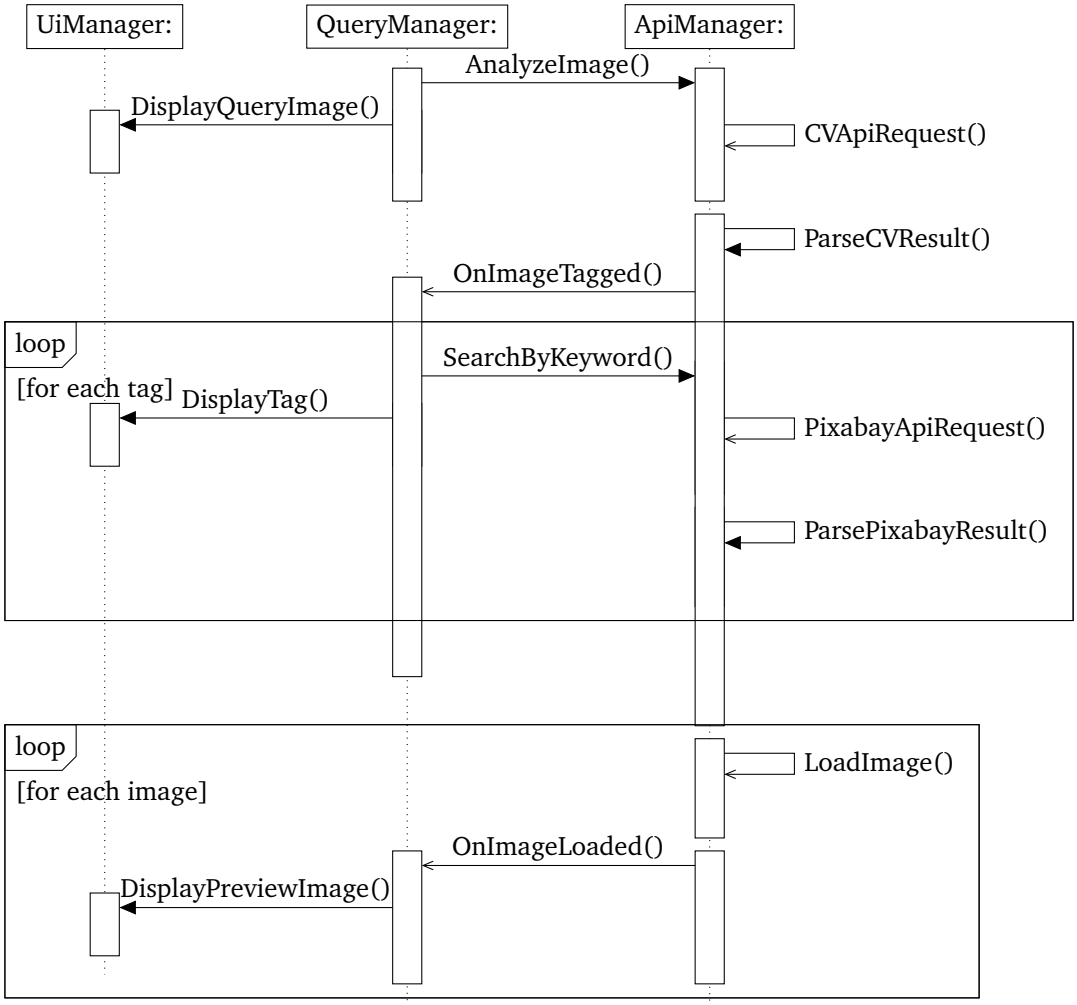


Fig. 4.9: Sequence of query initiation (parameters are omitted for reasons of clarity)

4.2 Developer and User Manual

This section gives a short overview of the setup and usage of the implemented application. Section 4.2.1 explains the required tools and processes for deployment while section 4.2.2 elaborates on the possible user interactions of the prototype.

4.2.1 Setup of the Prototype

In order to setup and start the implemented prototype, different tools have to be installed and executed. The following sections outline the required tools and the process of building and deploying the application.

Required Tools

The three parts essential for development and deployment of the prototype are the *Unity* editor, Microsoft's *Visual Studio* and a *HoloLens* device or emulator.

The application has been designed using the HoloLens Technical Preview (HTP) based off Unity 5.4.0f3. A newer version of Unity has been tried once during the development process but provoked an error, which is why Unity 5.4.0 has been used during the whole process of development (see section 4.1.1 for details). However for further development, the usage of the latest version of Unity is recommended.
Visual Studio 2015 Update 3 has been used as an integrated development environment (IDE) to develop and deploy the application on the HoloLens. For further development, the usage of the latest version of Visual Studio is possible.
Version 10.0.14393.0 of the HoloLens emulator has been used in combination with the Hyper-V feature of Windows 10 Education. The usage of Windows 10 Enterprise or Professional is possible as well.

Building and Deployment

In order to build and deploy the application, a series of steps has to be undertaken. After opening Unity, the provided project can be opened by selecting the project folder in the corresponding dialog. The project can then be built by choosing the *Build settings* dialog under *File* and then clicking on build and choosing the *App* folder inside of the project. After the building process is completed, the corresponding solution in the App folder can be opened in Visual Studio. After choosing the *Release* solution configuration and the *x86* solution platform, there are three possibilities of deploying the application:

- Deploying to the emulator by choosing it as the build target
- Deploying via Universal Serial Bus (USB) by choosing *Device* and pairing the device
- Deploying via wireless local area network (WLAN) by choosing *Remote Machine* and supplying the device's IP address

4.2.2 Usage of the Prototype

After the application started and the Unity splash screen has disappeared, the application prompts the user to place the result canvas (see the corresponding section below). After the initial placement of the result canvas, the application offers different interaction possibilities. Each of these interactions is executed by performing the HoloLens air tap gesture.
The most basic interaction is the initiation of a query by gazing at a scene and performing an air tap gesture. The possibility of this interaction is indicated by the appearance of a camera icon in the middle of the view. After a photograph of the scene has been taken, it is displayed as a query UI situated in the scene.
Every interaction apart from the query initiation is done by tapping on images or buttons displayed in the situated tag matrix or the result canvas.

Query UI

The query UI consists of a query image with two buttons in the middle and different tags with preview images placed all around. The following interactions are possible on the query UI:

- tap on query image: minimize or maximize tag matrix
- tap on tag: select tag as keyword for searching
- tap on preview image: select the chosen image as query image
- tap on arrow: revert to the previous query image
- tap on cross: delete query

Deleting the query only removes the situated photograph with the tag matrix. Tags that have been selected previously remain selected and still contribute to the result on the result canvas.

Result Canvas

The only interactive part of the result canvas is its header. The user can tap on the header to reset all keywords and clear the result. If there are no keywords to be cleared, a tap on the header starts the result placement mode. Placing the result canvas is done by gazing at the desired place and performing an air tap gesture. If there is a spatial mapping mesh available, the application automatically adjusts the position of the result canvas to fit the surface the user is gazing at. Shortly after starting the application, it might take a while for the mapping of the room to be created. When no mesh is available, the canvas is placed at a certain distance in front of the user.

4.3 Discussion of the Prototype

The implementation realizes the basic concept of situated photograph queries in order to provide a prototype for the evaluation in chapter 5. It incorporates the usage of a photograph metaphor to extract information about the user's surroundings and situates matrices of tags and images in the scene. A separate view is used to display the result of the query.
However there are also important differences between the original concept and the implemented prototype. The individual differences are summarized and explained in table 4.1. The three main reasons for these discrepancies named in the table can be explained as follows: *Complexity* describes the wish to keep the prototype simple in order to evaluate the basic concept. *Time* signifies the limited development time determined by the schedule for the present work. *Technical constraints* of the HoloLens device impeded the implementation of certain concepts.

Tab. 4.1: Differences between concept and implementation

Difference	Explanation	Reason
No search for visually similar images has been implemented.	As a substitute, browsing through preview images has been enabled instead.	complexity
The result canvas is not connected to another device.	No connection was necessary for a general evaluation of the query specification.	complexity
No zooming and indication of the photographed detail has been implemented.	The indication is difficult on the HoloLens because it has a limited field of view.	time, technical constraints
No combined tags in the tag matrix have been implemented.	The combination of tags on the result canvas has been implemented as a substitute.	complexity, time
The dynamic exploration of the tag matrix has not been implemented.	The minimization of the tag matrix has been implemented instead.	time
No search filters have been implemented.	Colors have been treated as tags and can be used as well.	complexity, time
No natural gestures have been used during the implementation and buttons were used.	The HoloLens only supports a general gesture in different variations.	technical constraints

Based on the assessment of the differences between prototype and concepts, varying kinds of extensions emerge: As proposed in the original concept, the tag matrix can be enriched with different filters. The combination of tags and filters can be implemented as well as the search for visually similar images. The extended tag matrix could be made dynamic to allow further ways of browsing. Other possible extensions that were not part of the original concept include a visualization of the confidence of different tags or actual interaction with retrieved images. For a more advanced prototype, performance related work regarding memory usage and performance bottlenecks should also be done. Altogether the prototype offers a solid basis for different possible extensions into diverse directions.

5 Evaluation of Prototype and Concept by User Study

The prototype described in chapter 4 has been implemented in order to evaluate the basic concept of situated photograph queries. Because of the novelty of the proposed concept, two predeterminations have been made:

1. Usability testing should be performed in order to verify the general applicability of the concept.
2. In such an early state, the prototype should not be compared quantitatively against established search interfaces.

These two predeterminations acted as a basis for the design of the user study explained in section 5.1. The results of the study are later presented and discussed in section 5.2.

5.1 Design of the User Study

The novelty of the HoloLens could potentially distort the evaluation of the concept and prototype: When using the device for the first time, problems with the usage and the fascination for AR based on HMDs could affect the study either way. Therefore seven participants were chosen that were planning to work with the HoloLens and thus had already received an introduction to the device before the study took part. The user study consists of two main parts: The usability testing that is conducted with the application (see section 5.1.1) and the questionnaire that is completed by the participants directly afterwards (see section 5.1.2).

5.1.1 Usability Testing

Usability testing has been performed in an experimental setting in one of the laboratories of the Chair of Multimedia-Technology. A small pilot study had been conducted before with two members of the staff of the chair in order to identify weaknesses of the usability testing. This study had shown that the planned introduction to the application during usability testing was too similar to the first tasks carried out after the introduction. Consequently, for the main study the users were given the introduction to the application using screenshots on a desktop computer before usability testing in order to save time and avoid the duplication of interactions.

Although the goal of the pilot study was to refine the design of the usability testing, the two participants already showed different approaches when using the application: The participant who has more experience with the HoloLens device and is an expert in AR took photographs from a smaller distance, got more correct tags and used less image browsing. Contrastingly, the participant who is an expert in information retrieval and has less experience with the HoloLens device took photos from further away, got more incomplete tags and used more image browsing.

The final process of usability testing contained the following three main steps:

1. Preparing the room to provide different interaction possibilities.
2. Assigning different tasks to the users to provoke interaction.
3. Taking measurements while users are interacting with the application.

These steps of usability testing are explained in the next three sections.

Preparation of the Room

Fig. 5.1: Photograph of the laboratory used for the study

The laboratory of the Chair of Multimedia-Technology contains several tabletops, desks and chairs (see figure 5.1). The desks were filled with a printer, a router, a telephone, remote controls, displays, mice and keyboards. This office setting was enriched with various objects that were placed inside the room to provide different possibilities for interaction (see figure 5.2). These objects included a green cap, three

tennis balls, a bunch of bananas and three apples, a yellow cup, a pair of scissors, an envelope with a cat photograph inside, a calendar with photographs of gardens and a folder with a picture of a giraffe on it.

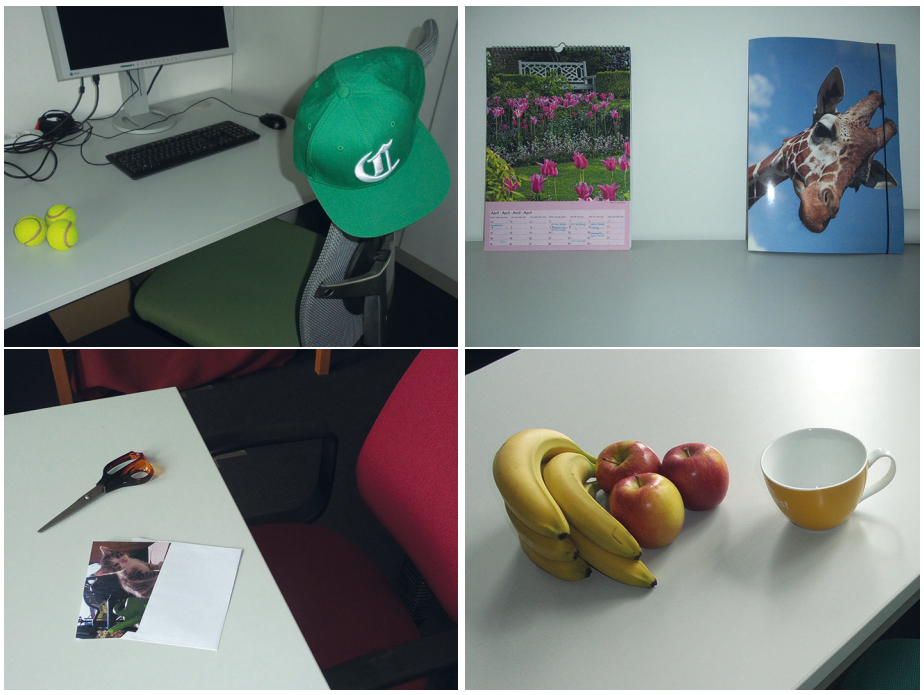


Fig. 5.2: Objects placed in the room

Tasks assigned to the Participants

In order to provoke interaction between the user and the application, a series of tasks had been elaborated. Every task consisted of one or two keywords that the user should search for. The task was considered as completed whenever the keyword or keywords were active on the result canvas. Simple tasks were set at the beginning in order to facilitate getting started.
The asked keywords were: *cup*, *giraffe*, *sport*, *apple banana*, *green banana*, *animal grass*, *indoor plant*, *indoor tennis*, *outdoor office*, *remote printer* and *garden scissors*. Due to the varying performance of the garden recognition, only three of the seven participants were able to complete this task. However because the actual completion of tasks was not the focus of the study, this fact did not cause any problems. Indeed the opposite was the case: The completion of the task not being trivial made the participants reveal their strategy.

Measurements during Usability Testing

Two different approaches have been applied to perform measurements during usability testing: The gathering of quantitative data by data logging and collecting qualitative data through user statements. Additionally, observations that have been made either monitoring the data logging or watching the participant were noted using pen and paper.

Quantitative logging of activities: A logging of activities was implemented in order to follow the participants' actions during usability testing. A log message was created for each of the following user activities:

- taking a photograph
- resetting the result
- activating or deactivating the result placement mode
- activating or deactivating a keyword
- tapping a preview image for browsing
- tapping the undo or the delete button
- minimizing or maximizing the tag matrix

Additionally the system triggered a log message whenever an image received a new tag and when the result changed. All log messages were recorded with a timestamp.

Qualitative statements of the users: Users were asked to give statements about the application whenever they encounter a special situation or problem. These statements were recorded using a camcorder that unobtrusively sat in one corner of the room.

5.1.2 Questionnaire

The questionnaire prepared for the user study had been divided into two parts: Part one was completed before usability testing and included general questions about the participants' age, gender, uncorrected problems with vision and previous experiences with AR, HMDs and the HoloLens.

The second part of the questionnaire was completed after usability testing in order to collect data about the users assessment of the application. The questionnaire was grouped into *pleasure*, *usefulness*, *effort*, *visual appearance* and the *general application*. This kind of grouping was chosen over a grouping into different parts of the application in order to facilitate comparison: It is easier for the participant to compare assessments like pleasure for different parts of the application when he gives all pleasure-related answers at once.

Each of the areas consisted of statements with 5-point Likert scales and questions with free text fields. This combination was chosen in order to gather comparable data and also allow remarks that have not been part of the statements before.

Pleasure, Usefulness and Effort

As defined by the International Organization for Standardization (ISO) in their standard ISO 9241, Usability marks the *effectiveness*, *efficiency* and *satisfaction* when using a system. In order to query these three factors in the questionnaire, representative feelings have been chosen:

1. *Pleasure* to signalize satisfaction
2. *Usefulness* to signalize effectiveness
3. *Effort* to signalize efficiency

Each of these factors has been polled from the participants using a 5-point Likert scaling and statements about different parts of the application (see table 5.1). Additionally, two free text fields asking for parts that were especially unpleasant/useless/cumbersome to use and especially pleasant/useful/effortless to use respectively were provided.

Tab. 5.1: Statements used to poll pleasure, usefulness and effort

Pleasure from *unpleasant (1)* to *pleasant (5)* • photographing scenes to extract tags • choosing keywords from the tag matrix • browsing through pictures in the tag matrix • the whole process of getting the right keywords
Usefulness from *not useful at all (1)* to *very useful (5)* • photograph cursor • placement of photographs in the room • connection between photographs and tags • preview images under each tag • browsing through preview images in the tag matrix
Effort from *effortless (1)* to *cumbersome (5)* • photographing scenes to extract desired tags • selecting keywords from the tag matrix • browsing through pictures in the tag matrix • the whole process of getting the right keywords • resetting the result

Visual Appearance

In addition to the three factors described above, the visual appearance of the application has been polled in order to determine if the scaling and positioning of elements is appropriate.
Besides one free text field asking for remarks about the visual appearance of the application, the following specific statements about different parts of the application were used in conjunction with a 5-point Likert scale:

- The **size** of the photo cursor, the photograph, the tag matrix and the preview images ranging from *too small (1)* to *too big (5)*.
- The **placement** of photo cursor and photograph ranging from *too near (1)* to *too far (5)*.
- The **number** of preview images ranging from *too small (1)* to *too big (5)*.
- The photographed **detail** ranging from *too narrow (1)* to *too wide (5)*.

General Application

The last part of the questionnaire concerns the application in general and utilizes the well-established *System Usability Scale (SUS)* proposed by John Brooke in 1996 [BO96]. It consists of the ten statements with 5-point Likert scales, whereby 1 means strongly disagree and 5 means strongly agree. The statements are alternating in their connotation so that the participants have to think about the meaning of each statement in detail before giving a rating [BO96].
Besides a free text field asking for remarks about the application in general, the following statements were used to assess the global usability of the application:

- I think that I would like to use this application frequently.
- I found the application unnecessarily complex.
- I thought the application was easy to use.
- I think that I would need the support of a technical person to be able to use this application.
- I found the various functions in this application were well integrated.
- I thought there was too much inconsistency in this application.
- I would imagine that most people would learn to use this application very quickly.
- I found the application very cumbersome to use.
- I felt very confident using the application.
- I needed to learn a lot of things before I could get going with this application.

5.2 Results

Seven male students aged between 22 and 26 years old (average 23.4) took part in the study. One participant reported to have a monocular blur, while all the others did not report any uncorrected problems with their vision. One of the participants had used the HoloLens for more than a short test, all the others tested the device once. None of the participants reported to have more experience with HMDs in general compared to the experience with the HoloLens. One participant stated that he had not experienced AR before, all the others had experienced AR once or a few times.

The two opposing ways of interacting with the prototype that emerged during the pilot study could also be observed during the usability testing. In order to confirm these observations, the first step of data analysis was to investigate the user behavior by examining the data logged during the usability testing.
After the data confirmed the existence of two different interaction approaches, the second step of analysis involved the inspection of the ratings given by the users through the Likert scales in the questionnaire.
The third category of data that has been investigated were the free text answers of the questionnaire and remarks made by the users during the usability testing.
The fourth and last data analyzed were the observations during the usability testing that concerned the user behavior in the room and different search strategies applied.

5.2.1 Logging of User Behavior

As a starting point for analysis, the total amount of time spent using the application and the total number of interactions carried out by the user has been counted for each participant. However these absolute measurements heavily depend on the performance of the Computer Vision API. Therefore the approach for analyzing the logged data was to calculate relative measures that are independent from external factors.

Differences in Browsing Rate

With regard to the observations made in the pilot study and during usability testing, the aim of the analysis of the logged data was to search for differences in the browsing behavior of the participants. To investigate this, the number of browsing interactions (i.e. the user tapping on a preview image) has been counted and divided by the total amount of user interactions for each of the participants. In this way, the *browsing rate* parameter that indicates the percentage of browsing during the interaction could be collected and compared between the participants.
The data that has been gathered in this way supports the initial observation: Four

of the participants had browsing rates of 8.9%, 8.3%, 7.0% and 0.7% respectively (average 6.2%) while the other three had browsing rates of 23.2%, 17.6% and 14.3% (average 18.4%). In other words, out of the 151 browsing interactions counted in total, 121 (80.1%) were executed by three of the seven participants and only 30 (19.9%) by the other four. The group of participants that preferred browsing are called browsing participants, with the other group being referred to as non-browsing participants.

Differences between Browsing and Non-Browsing Participants

In the same way the browsing rate has been calculated, the *photograph rate* has been gathered as additional parameter that indicates the percentage of photographing interactions.
While not being significantly different, the photograph rate is slightly lower for all browsing participants (29.9%, 26.2% and 26.2%: average 27.4%) compared to each of the non-browsing participants (35.3%, 32.6%, 31.7% and 31.3%: average 32.7%). One possible explanation for this difference is that a non-browsing participant takes a new photograph whenever the tags of the previous query are not sufficient, while a browsing participant starts to browse through images of the query and therefore does not need a new photograph.
Another comparable relative measure is the rate of user interactions per minute. This parameter is obtained by dividing the total number of interactions by the total time spent using the application. The average user interactions per minute for non-browsing participants (4.8 interactions per minute) is lower than for browsing participants (5.4 interactions per minute). This possibly results from the fact that browsing often incorporates rather quick jumping from one picture to the next, which incorporates comparatively many tapping interactions in a short period of time.
Although absolute measures should be interpreted very cautiously, it is worth noting that the average time spent using the application is higher for browsing participants (38 minutes and 36 seconds) compared to non-browsing participants (26 minutes and 18 seconds) and the average number of total interactions per user is also much higher for browsing participants (209) compared to non-browsing participants (124). This could be an indicator that browsing is more time consuming an involves more interactions compared to a non-browsing behavior.

Differences in Photograph Delete Rate

In the same way that the browsing rate and the photograph rate have been obtained, the *delete rate* can be retrieved and signifies the percentage of delete operations compared to the total number of operations executed by the user. Delete rates on

their own fluctuate for all participants between 16.8% and 31.3% with an average of 22.4%. There are no significant differences between browsing participants (average 21.8%) and non-browsing participants (average 22.9%).
However another interesting parameter can be retrieved from this data: When dividing the delete rate by the photograph rate, one can obtain the *photograph delete rate*. This parameter signifies the percentage of queries that the participant deleted. A photograph delete rate of 100% means the participant eventually deleted every query that he created, while a photograph delete rate of 0% means that the participant deleted none of the queries he created. In other words: Participants with lower photograph delete rates rather tend to keep queries instead of deleting them. Indeed there are significant differences between photograph delete rates: Three participants exhibit a photograph delete rate of 95.7%, 91.5% and 85.5% respectively (average 90.9%) while the other four only have 68.8%, 63.6%, 61.1% and 53.1% (average 61.7%). The latter group that tends to keep queries is referred to as the keeping participants while the former are called non-keeping participants.

Differences between Keeping Participants and Non-Keeping Participants

When comparing participants' types with regard to browsing and keeping queries, it is salient that all but one browsing participants are non-keeping participants and all but one non-browsing participants are keeping participants (see table 5.2). This observation could explain why the photograph rate is not significantly lower for browsing participants: Even though browsing participants should need fewer photographs in the first place, they also tended to delete their queries more often than non-browsing participants and therefore needed to take new photographs more often, which levels out the photograph rate compared to non-browsing participants.

Tab. 5.2: Different participants and groups assigned to them

Participant number:	1	2	3	4	5	6	7
Keeping participant (+) or not (-)	-	+	-	+	-	+	+
Browsing participant (+) or not (-)	-	-	+	-	+	+	-

When investigating absolute measures of both groups, a big difference in the average time spent using the application can be seen: While keeping participants spent an average time of 25 minutes and 3 seconds with the application, non-keeping participants used it 40 minutes and 16 seconds. One explanation for this could be that keeping participants can use keywords from older queries while non-keeping participants need to retrieve them once again.

5.2.2 Rating through Likert Scales

The rating the users conveyed using the Likert scaling showed significant differences between browsing participants and non-browsing participants in some parts. Because of the overlap of non-keeping participants with browsing participants and keeping participants with non-browsing participants, the data for these overlapping groups is generally quite similar. Therefore keeping and non-keeping participants are mentioned especially whenever the data differs from the non-browsing and browsing participants respectively.

Pleasure

Analysis of the Likert scales revealed that choosing keywords from the tag matrix was slightly more pleasurable than the other interactions (see table 5.3). When comparing different groups of participants, it is salient that in average browsing participants stated lower pleasure for all interactions compared to non-browsing participants (see figure 5.3). Interestingly, this difference is also visible between keeping participants and non-keeping participants for photographing scenes and choosing keywords, but not for browsing through pictures and the whole process.

Tab. 5.3: User ratings for the pleasure felt when using the application

Participant number:	1	2	3	4	5	6	7
Keeping participant (+) or not (-):	-	+	-	+	-	+	+
Browsing participant (+) or not (-):	-	-	+	-	+	+	-
1. Photographing scenes to extract desired tags:	4	5	2	3	4	4	5
2. Choosing keywords from the tag matrix:	4	4	3	5	4	4	5
3. Browsing through pictures in the tag matrix:	4	3	4	5	4	2	5
4. Whole process of getting the right keywords:	4	4	4	4	3	3	5

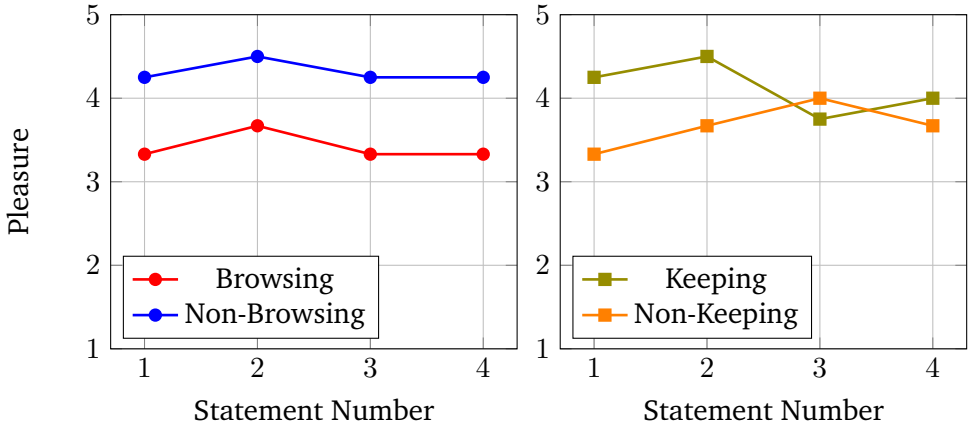


Fig. 5.3: Comparison of pleasure between groups of participants

Effort

The given ratings show that photographing the scenes and the whole process of finding the right keyword were requiring the most effort (see table 5.4). These two statements also show the biggest difference between browsing participants and non-browsing participants: Browsing participants found both much more cumbersome (see figure 5.4). One explanation for this is that participants who found photographing the scene difficult were applying browsing to achieve their goal, which was less efficient and therefore made the whole process more cumbersome.
Keeping participants and non-keeping participants do not exhibit such a big difference in these statements. Instead, keeping participants found the selection of keywords from the tag matrix much more effortless than non-keeping participants. This can be explained by the fact that they could choose keywords from older queries when possible, which requires less effort than taking a new photograph.

Tab. 5.4: User ratings for the effort needed while using the application

Participant number:	1	2	3	4	5	6	7
Keeping participant (+) or not (-):	-	+	-	+	-	+	+
Browsing participant (+) or not (-):	-	-	+	-	+	+	-
1. Photographing scenes to extract desired tags:	2	2	3	2	3	3	2
2. Selecting keywords from the tag matrix:	2	2	2	1	2	1	1
3. Browsing through pictures in the tag matrix:	2	2	1	1	3	2	1
4. Whole process of getting the right keywords:	2	2	4	3	4	3	2
5. Resetting the result:	1	1	1	1	1	1	2

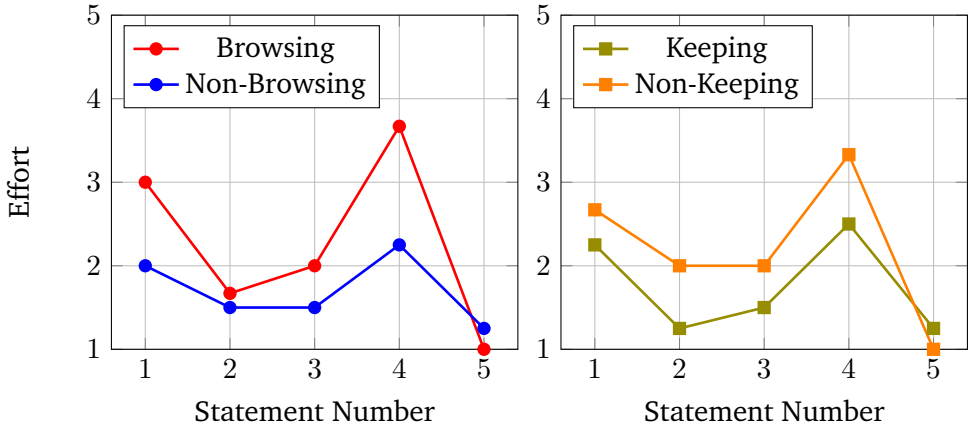


Fig. 5.4: Comparison of needed effort between different groups of participants

Usefulness

The examination of the ratings for usefulness revealed that they are widely spread for all statements (see table 5.5). However when comparing the different groups of

participants, a clearer picture can be perceived (see figure 5.5).
Non-browsing participants found the placement of photographs in the room more useful than non-browsing participants. In return, browsing participants found browsing through pictures more useful than non-browsing participants.
Both these differences can also be found between keeping participants and non-keeping participants. Additionally, keeping participants found the connection between photographs and tags more useful than non-keeping participants, while the latter found the preview images under each tag more useful than the former.
This can be an indicator that participants who tend to keep queries have to remember where the tags were and therefore mentally connect tags with the scene they have photographed. The preview images are less useful in this process.
In contrast, participants who do not keep queries and take a photograph each time evaluate the tags by looking at the preview images, because this requires less mental effort than reading the tags.

Tab. 5.5: User ratings for the usefulness of the application

Participant number:	1	2	3	4	5	6	7
Keeping participant (+) or not (-):	-	+	-	+	-	+	+
Browsing participant (+) or not (-):	-	-	+	-	+	+	-
1. Photograph cursor:	3	3	5	4	4	2	5
2. Placement of photographs in the room:	4	4	3	5	2	4	5
3. Connection between photographs and tags:	4	5	2	3	4	5	5
4. Preview images under each tag:	5	3	3	4	5	2	3
5. Browsing through pictures in the tag matrix:	4	4	5	3	5	3	3

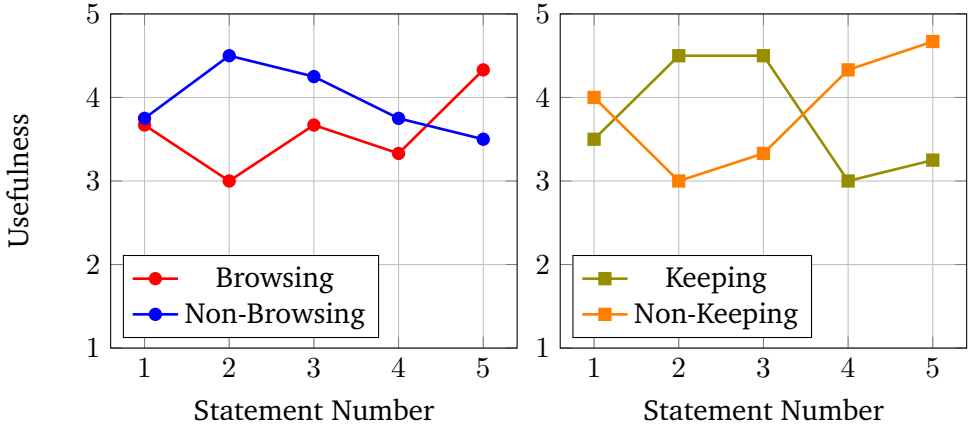


Fig. 5.5: Comparison of usefulness between different groups of participants

Visual Appearance

While most of the visual appearance did not receive extreme ratings, it is salient that the values are spreading a lot for the size of the photographed detail (see table

5.6). This goes hand in hand with the fact that some participants remarked that the photographed detail of the scene was sometimes too wide and sometimes too small. Another statement with spreading values is the number of preview images: Browsing participants would have liked more images while non-browsing participants would prefer fewer of them. This is also intuitively explainable, as more preview images provide a broader possibility for browsing.

Tab. 5.6: User ratings for the visual appearance of the application

Participant number:	1	2	3	4	5	6	7
Keeping participant (+) or not (-):	-	+	-	+	-	+	+
Browsing participant (+) or not (-):	-	-	+	-	+	+	-
The size of the photo cursor was...	3	3	3	3	3	3	3
The placement of the photo cursor was...	3	3	2	2	3	3	3
The photographed detail of the scene was...	2	3	1	2	4	3	3
The placement of the photograph was...	3	2	2	2	2	3	3
The size of the photograph was...	3	2	3	2	3	3	3
The size of the tag matrix was...	3	3	3	3	3	2	4
The size of the preview images was...	3	3	3	3	3	3	3
The number of preview images was...	3	2	2	4	2	2	4

System Usability Scale Score

In order to obtain the SUS score, each rating has to be normalized so that it ranges from 0 (worst) to 4 (best). These normalized ratings are then added together and multiplied by 2.5 to obtain the final score (see table 5.7). Although scores range from 0 to 100, they should not be considered as a percentage. Research has shown that a score of 68 can be considered to be average [BO96].

The average total score achieved by the present application is 76.429. The total score averages 72.5 for browsing participants and 79.375 for non-browsing participants. Although differences between the groups of participants are generally not big, it is worth noting that browsing participants rated the application lower in 7 of the 10 statements compared to non-browsing participants. The biggest difference can be seen in the confidence of the users, where browsing participants and non-keeping participants rated lower on average than non-browsing participants and keeping participants respectively. Another difference is visible in the complexity rating, which is on average worse from browsing participants than from non-browsing participants.

Tab. 5.7: System Usability Scale ratings from each participant

Participant number:	1	2	3	4	5	6	7
Keeping (+) or not (-):	-	+	-	+	-	+	+
Browsing (+) or not (-):	-	-	+	-	+	+	-
Frequency of usage	3	1	3	2	0	3	1
Complexity	3	4	4	2	1	2	4
Ease of use	3	3	4	3	4	1	4
Need of support	3	4	4	4	3	3	4
Integration of functions	2	3	3	3	3	2	4
Consistency	4	3	3	4	4	3	4
Ease of learning	3	4	4	3	4	3	4
Effort of use	3	3	3	3	3	3	4
Confidence	3	3	1	2	2	3	3
Previous knowledge needed	2	4	4	4	4	3	4
Total SUS score	**72.5**	**80.0**	**82.5**	**75.0**	**70.0**	**65.0**	**90.0**

5.2.3 Free Text Answers and Remarks during the Study

In order to facilitate the analysis of the qualitative data, free text answers from the questionnaire have been consolidated with the remarks made during the study and then categorized into five groups. They either concerned the *photographed detail of the scene*, the *tap gesture and buttons*, *overview and context* of the situated queries, *tags and browsing* or the *performance* of the application.

Photographed Detail of the Scene

Four of the participants commented that they would like to have an indicator for the photographed detail of the scene prior to taking the photograph. Two of them noted that the detail of the photograph is too narrow in some cases and too wide in others. One suggested that a zoom function would be suitable to overcome this problem. This shows that this function, which is a part of the original concept, should be one of the next extensions for the prototype.

Tap Gesture and Buttons

Four of the participants encountered problems when hitting a button, three of them also had problems with the recognition of the air tap gesture. These two problems might be connected to each other: If a gesture is not recognized when tapping on a button, it feels like hitting the button is difficult. This is an indicator that even more training should be done with the HoloLens before a user study in order to avoid such

problems.
Another participant had his finger in the picture and asked for a possibility to cut it out of the photograph or to get some feedback when the finger is actually in the recognition zone of the HoloLens. This could be realized by displaying the photograph cursor only when the hand is recognized by the system.

Overview and Context

Different remarks have been made with regard to the overview and context of the queries. One participant stated that saving pictures in the room is a useful idea. One participant asked about an overview map or context menu with the number of queries situated in the room in order to get an overview and to find the queries. He also asked about the possibility to restore deleted queries. Such possibilities to gain an overview have not yet been a part of either the concept or the prototype, but should be a point for future development.
Two participants encountered an unexpected minimization of the tag matrix, while two others had problems with overlapping matrices. One of them encountered a problem with the visibility of the matrix at a certain angle. Further testing is needed in order to determine what caused this unexpected behavior.
Another participant asked about the possibility to move queries, because his query was situated against the light and therefore difficult to see. Two other participants also reported problems with visibility against the light. While problems with lighting do not concern the concept or prototype, it is arguable if a function to move queries should be incorporated.

Tags and Browsing

A number of remarks also concerned tags and browsing, which is the core part of the application. Three of the participants would have liked to use combined tags to browse through images. Although browsing is not included in the original concept, a draft for combining tags inside of the tag matrix has been proposed. This could be combined with a browsing function in order to achieve the desired feature.
Three other participants wanted more preview images for browsing, another participant wanted to browse into a category without having to select a particular image. A varying number of preview images like proposed in the original concept would depict such a feature: When approaching a tag, more preview images appear, which is the equivalent for browsing into that category.
One participant asked if tags are combined over time when browsing through images. This remark is evidence to suggest that some people expect to gather information like described by the berry picking model of information seeking (see section 2.1.3).
One participant stated that extracting keywords was easy when the tags met the

expectation, but got frustrating when they did not. Another participant asked for further tags without images. While images have been part of the concept all along, a growing number of tags has also been proposed in the original concept. One wanted for a loading indicator while the system is searching for tags, which is a reasonable suggestion.

Performance

The last category of remarks has been made regarding the performance of the application. This concerned the performance of external dependencies like the API or the HoloLens as well as internal problems. Users reported a lag when lots of images were downloaded and the application crashing during long sessions. Further investigation is needed in order to understand and eliminate these problems.

5.2.4 Observations during the Study

The last category of measurements are the observations that were made during the process of the usability testing. The purpose of this qualitative and subjective data is to discover user behavior that is not visible in the logged data. The observations have been divided into two categories: The behavior of the user in the room when interacting with the application and the searching strategies that have been applied by the users.

Behavior in the Room

The behavior in the room varied considerable between the groups of participants. Two of the browsing participants were standing comparably far away from the objects when taking a photograph, which resulted in less precise tags. In contrast, three of the non-browsing participants improved their tags by adjusting their position after a failed try.
Non-browsing participants also rearranged the objects more often: One did this even before he tried a first photograph, another one always tried to isolate the objects completely. To achieve this, he was the only one to put other objects away and hold objects in front of himself while photographing. Another non-browsing participant even turned the calendar pages, which no one else did. He also asked if it was possible to draw something on the whiteboard during the study. Out of the browsing participants, only one adopted rearranging objects later in the course of the study.
One of the browsing participants was quite unobservant and only spotted the giraffe after five minutes. Although general implications should not be made, it is salient that browsing participants were less aware of the room and used it less during their

interaction with the application. Further research has to be done to investigate this phenomena more profoundly.

Searching Strategies

The searching strategies applied by the users also match their groups. Two of the non-keeping participants found *sport* through a picture of bowling balls under the *indoor* tag rather than by photographing the tennis balls. In contrast, two of the keeping participants saved a lot of time and effort because they were able to select *indoor tennis* completely from previous queries.
Browsing participants seem to take more spontaneous decisions instead of prior plans: One found *indoor plant* by photographing the room, selecting the *indoor* tag and then browsing to a picture of wood under the *floor* tag. Another one got the *outdoor* tag by browsing through images of buildings under an *indoor* tag. He also got *green* by browsing through images of apple trees in the photo of the banana.

5.2.5 Discussion of Results

Seven participants took part in the user study which had been composed of a usability testing and a successive questionnaire. The usability testing involved a total 3 hours and 41 minutes of interaction time with the application. Overall 224 ratings have been gathered using 5-point Likert scales with different statements.

Findings of the User Study

After analyzing the logged data for conspicuousness, two salient ways of grouping the participants emerged based on the tendency to browse through images and the tendency to delete previously used queries.
Although the number of participants does not allow any statistical inferences, a number of findings have been made that give a reasonable impression. In summary, participants who applied a browsing approach also wanted more extensive browsing features. They needed more effort and felt less pleasure when using the application. Participants who did not apply this approach found the placement of photographs in the room much more useful.
Participants who used to keep older queries found the connection between photographs and tags more useful and were solving the tasks faster. They found the preview images much less useful.
Overall, these findings indicate that the current prototype suits participants that tend to keep older queries and do not apply a browsing approach during the search.

Suggestions for the Future

The present prototype is only a first basic implementation of the concept of situated photograph queries. In order to produce a more complemented experience of image retrieval in AR, several improvements should be made in the future. Based on the findings of the user study, the following suggestions are made concerning the prototype:

1. **Implement parts of the original concept** like combined tags, dynamic exploration of tags and preview images and zooming the photographed detail of the scene.
2. **Think about additional concepts** to provide the users with more overview and context for the situated queries.
3. **Improve performance and user experience** by investigating system crashes and giving more feedback of the systems state.
4. **Tailor the application for different groups** for example by providing further tags without pictures and allowing more extensible browsing.

Additionally, further user studies should be made in order to investigate possible correlations between spatial awareness and applied searching strategies more profoundly.

6 Conclusion

The main goal of the present work was to elaborate on the possibility of improving the process of image retrieval using augmented reality. The approach of the present work is being summarized in section 6.1 while section 6.2 gives an outlook on future implications that can be derived from the thesis.

6.1 Summary of the Present Work

It was illustrated in chapter 1 that a conjunction of image retrieval with augmented reality has potential. Head-mounted displays are currently on the rise with several devices introduced during the last years. Theses kind of devices imply and provide a completely new way of interacting with data. Such a new form of interaction provides the possibility for image retrieval to adopt to rising numbers of pictures and changing needs of people.
A general trend towards natural search interfaces as assessed by Marti Hearst can be achieved using augmented reality [Hea11]. Additionally, the connection of reality and virtuality can provide a direct way of retrieval whenever an information need is triggered by the surroundings. This can lead to a convenient interaction by making use of familiar objects and locations.

To evaluate the usefulness of augmented reality for image retrieval, a concept was needed that could be implemented and tested with users. This concept in turn should build upon related work and meet some requirements.
In order to determine requirements for image retrieval within augmented reality, definitions, classifications and design challenges of both fields have been investigated in chapter 2. Section 2.1 has examined image retrieval in detail in order to define the term, investigate possibilities for classification and identify different design challenges. In a similar way, AR has been examined in section 2.2. Based on the findings of these sections, a taxonomy for image retrieval within AR has been proposed in section 2.3.
The taxonomy consists of two parts: One for session parameters that allows to describe general characteristic of user, context and data and one for the interaction process that consists of *natural query specification*, *situated result visualization* and *3D result interaction*.

In chapter 3 the interaction process part of the taxonomy was picked up in order to provide structure for related work and elaborated concepts. Section 3.1 examined

different research work that involves concepts relevant for potential usage in image retrieval within augmented reality.
After examining related work, a brainstorming for general interaction concepts has been conducted for each of the steps of the interaction process taxonomy. Ideas that emerged during that brainstorming have been presented in section 3.2.
The session part of the taxonomy has then been used to describe different application scenarios that act as a part of the requirements for the desired concepts in section 3.3. The other part of the requirements has been embodied in the three design goals of *novelty*, *variety* and *usability*. Each of these goals consists of several subgoals that have been deduced from the design challenges identified in chapter 2.
Based on the general interaction ideas and the requirements determined in section 3.3, two comprehensive concepts for image retrieval within augmented reality have been elaborated in section 3.4: *Tangible Query Workbench* and *Situated Photograph Queries*.

The prototype described in chapter 4 acts as a connection between the concept and the evaluation. A Microsoft HoloLens has been chosen for development because it offers a stable augmented reality experience. The practicability of the implementation of concepts on the HoloLens has been outlined and entailed the choice of the situated photograph queries for implementation. The resulting prototype has been developed in an agile and iterative way to allow a flexible timetable. It depicts an expandable subset of the original concept that implements the basic idea and has undergone some changes.

A user study consisting of a usability testing and a questionnaire has then been conducted with the help of the implemented prototype. The study has shown that although a number of suggestions for improvement came up, the concept was generally liked by the participants. Users asked for features of the original concept that were not implemented due to technical and time constraints, which shows that the design of the features was generally well-founded.
During data analysis of the user study, two different ways of grouping the participants based on their behavior emerged:

- Users were either employing a lot of image browsing or not.
- Users were either deleting older queries or not.

These groups of participants also showed characteristic behavior in the logged data and the questionnaire. Altogether, the evaluation provided interesting points of contact for further research in the field and revealed that the connection between image retrieval and augmented reality was well received by users.

6.2 Outlook on Further Work

The present work has provided several starting points for further research. The proposed taxonomy can be used as a basis for describing image retrieval sessions in augmented reality or conceiving new concepts for interaction in this field. The basic interaction ideas gathered in section 3.2 can act as inspiration for such concepts.
The comprehensive concept of a *tangible query workbench* can be refined and implemented in the future in order to provide a contrast to the already implemented concept of *situated photograph queries*. This prototypical implementation can be enriched with features that were in the original concept and new ideas in order to provide a more integrated user experience. Additionally, the concept can also be expanded into other domains like object recognition in tourism and museums or product search.
The findings of the conclusive evaluation of the prototype can be used as a starting point for a deeper investigation into the connection between spatial behavior and image retrieval strategies.

In summary, the present work provides a first insight into the possibility of practicing image retrieval within augmented reality. It has produced two comprehensive concepts, one of which has been implemented using a Microsoft HoloLens and evaluated during a user study. The study showed that the concept was generally well received by the users and gave interesting insights into the behavior of users in the room when fulfilling image retrieval tasks.

Bibliography

[Aar+10] Chris van Aart, Bob Wielinga, and Willem Robert van Hage. „Mobile Cultural Heritage Guide: Location-aware Semantic Search“. In: *Proceedings of the 17th International Conference on Knowledge Engineering and Management by the Masses*. EKAW'10. Berlin, Heidelberg: Springer-Verlag, 2010, pp. 257–271 (cit. on p. 37).

[AP00] Jürgen Assfalg and Pietro Pala. „Querying by Photographs: A VR Metaphor for Image Retrieval“. In: *IEEE MultiMedia* 7.1 (2000), pp. 52–59 (cit. on p. 34).

[Ass+02] J Assfalg, A Del Bimbo, and P Pala. „Three-dimensional interfaces for querying by example in content-based image retrieval“. In: *IEEE Transactions on Visualization and Computer Graphics* 8.4 (2002), pp. 305–318 (cit. on pp. 36, 45, 57).

[Ass+98] J Assfalg, A Del Bimbo, and P Pala. „Virtual Reality for Image Retrieval“. In: *Advances in Multimedia Information Systems: 4th International Workshop, MIS'98 Istanbul, Turkey September 24–26, 1998 Proceedings*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1998, pp. 199–204 (cit. on p. 35).

[Azu97] Ronald T Azuma. „A survey of augmented reality“. In: *Presence: Teleoperators and virtual environments* 6.4 (1997), pp. 355–385 (cit. on p. 16).

[Bar+16] C Barreiros, E Veas, and V Pammer-Schindler. „Pre-attentive Features in Natural Augmented Reality Visualizations“. In: *2016 IEEE International Symposium on Mixed and Augmented Reality (ISMAR-Adjunct)*. 2016, pp. 72–73 (cit. on pp. 38, 47).

[Bil+15] Mark Billinghurst, Adrian Clark, and Gun Lee. „A Survey of Augmented Reality“. In: *Found. Trends Hum.-Comput. Interact.* 8.2-3 (2015), pp. 73–272 (cit. on pp. 17, 19, 21).

[Bla+04] Alan F Blackwell, Mark Stringer, Eleanor F Toye, and Jennifer A Rode. „Tangible Interface for Collaborative Information Retrieval“. In: *CHI '04 Extended Abstracts on Human Factors in Computing Systems*. CHI EA '04. New York, NY, USA: ACM, 2004, pp. 1473–1476 (cit. on p. 36).

[BO96] John Brooke and Others. „SUS-A quick and dirty usability scale“. In: *Usability evaluation in industry* 189.194 (1996), pp. 4–7 (cit. on pp. 80, 87).

[Bow+04] Doug Bowman, Ernst Kruijff, Joseph J LaViola Jr, and Ivan Poupyrev. *3D User Interfaces: Theory and Practice, CourseSmart eTextbook*. Addison-Wesley, 2004 (cit. on p. 21).

[CC01] W Chen and Shih-Fu Chang. „VISMap: an interactive image/video retrieval system using visualization and concept maps“. In: *Image Processing, 2001. Proceedings. 2001 International Conference on*. Vol. 3. 2001, 588–591 vol.3 (cit. on pp. 36, 39, 46, 48).

[Che+05] Chufeng Chen, Michael Oakes, and Sharon McDonald. „Using a time and location combination clustering model for browsing personal images“. In: *Proceedings of British HCI*. Vol. 2. 2005, pp. 244–246 (cit. on p. 15).

[Che+06] Chufeng Chen, Michael Oakes, and John Tait. „Browsing Personal Images Using Episodic Memory (Time + Location)“. In: *Proceedings of the 28th European Conference on Advances in Information Retrieval*. ECIR'06. Berlin, Heidelberg: Springer-Verlag, 2006, pp. 362–372 (cit. on p. 15).

[Chi+05] Patrick Chiu, Andreas Girgensohn, Surapong Lertsithichai, Wolf Polak, and Frank Shipman. „MediaMetro: Browsing Multimedia Document Collections with a 3D City Metaphor“. In: *Proceedings of the 13th Annual ACM International Conference on Multimedia*. MULTIMEDIA '05. New York, NY, USA: ACM, 2005, pp. 213–214 (cit. on pp. 38, 47).

[Chr+10] Olivier Christmann, Noëlle Carbonell, and Simon Richir. „Visual Search in Dynamic 3D Visualisations of Unstructured Picture Collections“. In: *Interact. Comput.* 22.5 (2010), pp. 399–416 (cit. on p. 40).

[Cox+00] I J Cox, M L Miller, T P Minka, T V Papathomas, and P N Yianilos. „The Bayesian image retrieval system, PicHunter: theory, implementation, and psychophysical experiments“. In: *IEEE Transactions on Image Processing* 9.1 (2000), pp. 20–37 (cit. on p. 8).

[Dat+08] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z Wang. „Image Retrieval: Ideas, Influences, and Trends of the New Age“. In: *ACM Comput. Surv.* 40.2 (2008), 5:1–5:60 (cit. on pp. 5, 6, 8, 9, 14, 15, 24, 27–29).

[Eng+11] David Engel, Christian Herdtweck, Björn Browatzki, and Cristóbal Curio. „Image Retrieval with Semantic Sketches“. In: *Human-Computer Interaction – INTERACT 2011: 13th IFIP TC 13 International Conference, Lisbon, Portugal, September 5-9, 2011, Proceedings, Part I*. Ed. by Pedro Campos, Nicholas Graham, Joaquim Jorge, et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 412–425 (cit. on pp. 35, 45).

[Far+10] A Faro, D Giordano, C Pino, and C Spampinato. „Visual Attention for Implicit Relevance Feedback in a Content Based Image Retrieval“. In: *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications*. ETRA '10. New York, NY, USA: ACM, 2010, pp. 73–76 (cit. on pp. 41, 43).

[Fid97] Raya Fidel. „The image retrieval task: implications for the design and evaluation of image databases“. In: *New Review of Hypermedia and Multimedia* 3.1 (1997), pp. 181–199 (cit. on pp. 7, 23).

[Fri+15] O Fried, S DiVerdi, M Halber, E Sizikova, and A Finkelstein. „IsoMatch: Creating Informative Grid Layouts“. In: *Comput. Graph. Forum* 34.2 (2015), pp. 155–166 (cit. on p. 40).

[GI10] Ai Gomi and Takayuki Itoh. „MIAOW: A 3D Image Browser Applying a Location- and Time-based Hierarchical Data Visualization Technique“. In: *Proceedings of the International Conference on Advanced Visual Interfaces*. AVI '10. New York, NY, USA: ACM, 2010, pp. 225–232 (cit. on pp. 39, 48).

[Gir+11] B Girod, V Chandrasekhar, R Grzeszczuk, and Y A Reznik. „Mobile Visual Search: Architectures, Technologies, and the Emerging MPEG Standard“. In: *IEEE MultiMedia* 18.3 (2011), pp. 86–94 (cit. on pp. 36, 45).

[Gru+11] Jens Grubert, Tobias Langlotz, and Raphaël Grasset. „Augmented reality browser survey“. In: *Institute for computer graphics and vision, University of Technology Graz, technical report* 1101 (2011) (cit. on p. 37).

[Han97] Chris Hand. „A Survey of 3D Interaction Techniques“. In: *Computer Graphics Forum* 16.5 (1997), pp. 269–281 (cit. on p. 2).

[Har+97] Shouji Harada, Yukihiro Itoh, and Hiromasa Nakatani. „Interactive image retrieval by natural language“. In: *Optical Engineering* 36.12 (1997), pp. 3281–3287 (cit. on pp. 34, 41, 42, 44, 49).

[Har77] L R Harris. „User oriented data base query with the ROBOT natural language query system“. In: *International Journal of Man-Machine Studies* 9.6 (1977), pp. 697–713 (cit. on p. 34).

[Hea09] Marti A Hearst. *Search User Interfaces*. 1st. New York, NY, USA: Cambridge University Press, 2009 (cit. on pp. 2, 10–12, 16).

[Hea11] Marti A Hearst. „'Natural' Search User Interfaces“. In: *Commun. ACM* 54.11 (2011), pp. 60–67 (cit. on pp. 3, 26, 44, 50, 93).

[Hol+04] L Hollink, A.Th. Schreiber, B J Wielinga, and M Worring. „Classification of user image descriptions“. In: *International Journal of Human-Computer Studies* 61.5 (2004), pp. 601–626 (cit. on pp. 7–9, 12, 13, 15, 23–25).

[Hu+16] Ronghang Hu, Huazhe Xu, Marcus Rohrbach, et al. „Natural Language Object Retrieval“. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016 (cit. on pp. 41, 42, 49).

[Jes+07] Rui Jesus, Ricardo Dias, Rute Frias, and Nuno Correia. „Geographic Image Retrieval in Mobile Guides“. In: *Proceedings of the 4th ACM Workshop on Geographical Information Retrieval*. GIR '07. New York, NY, USA: ACM, 2007, pp. 37–38 (cit. on p. 37).

[Jet+11] Hans-Christian Jetter, Jens Gerken, Michael Zöllner, Harald Reiterer, and Natasa Milic-Frayling. „Materializing the Query with Facet-streams: A Hybrid Surface for Collaborative Search on Tabletops“. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '11. New York, NY, USA: ACM, 2011, pp. 3013–3022 (cit. on p. 36).

[Jia+06] Menglei Jia, Xin Fan, Xing Xie, Mingjing Li, and Wei-Ying Ma. „Photo-to-Search: Using Camera Phones to Inquire of the Surrounding World“. In: *7th International Conference on Mobile Data Management (MDM'06)*. 2006, p. 46 (cit. on pp. 36, 45).

[Jia+15] Meng Jian, Cheolkon Jung, Yanbo Shen, and Juan Liu. „Interactive image retrieval using constraints“. In: *Neurocomputing* 161 (2015), pp. 210–219 (cit. on pp. 41, 42, 49).

[Jos+98] Joemon M Jose, Jonathan Furner, and David J Harper. „Spatial Querying for Image Retrieval: A User-oriented Evaluation“. In: *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '98. New York, NY, USA: ACM, 1998, pp. 232–240 (cit. on pp. 35, 45).

[Kal+11] Denis Kalkofen, Christian Sandor, Sean White, and Dieter Schmalstieg. „Visualization techniques for augmented reality“. In: *Handbook of Augmented Reality*. Springer, 2011, pp. 65–98 (cit. on p. 20).

[Kap82] S.Jerrold Kaplan. „Cooperative responses from a portable natural language query system“. In: *Artificial Intelligence* 19.2 (1982), pp. 165–187 (cit. on p. 34).

[Kat+00] H Kato, M Billinghurst, I Poupyrev, K Imamoto, and K Tachibana. „Virtual object manipulation on a table-top AR environment“. In: *Proceedings IEEE and ACM International Symposium on Augmented Reality (ISAR 2000)*. 2000, pp. 111–119 (cit. on p. 19).

[KB02] BYOUNG CHUL KO and HYERAN BYUN. „Query-by-Gesture: An Alternative Content-Based Image Retrieval Query Scheme“. In: *Journal of Visual Languages & Computing* 13.4 (2002), pp. 375–390 (cit. on pp. 35, 45).

[KC13] Jin-Dong Kim and K Bretonnel Cohen. „Natural language query processing for SPARQL generation: A prototype system for SNOMED CT“. In: *Proceedings of biolink*. 2013, pp. 32–38 (cit. on pp. 34, 44).

[Kei+13] Jens Keil, Michael Zoellner, Timo Engelke, Folker Wientapper, and Michael Schmitt. „Controlling and Filtering Information Density with Spatial Interaction Techniques via Handheld Augmented Reality“. In: *Virtual Augmented and Mixed Reality. Designing and Developing Augmented and Virtual Environments: 5th International Conference, VAMR 2013, Held as Part of HCI International 2013, Las Vegas, NV, USA, July 21-26, 2013, Proceedings, Part I*. Ed. by Randall Shumaker. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 49–57 (cit. on pp. 41, 43, 48, 50).

[Klu+12] Stefanie Klum, Petra Isenberg, Ricardo Langner, Jean-Daniel Fekete, and Raimund Dachselt. „Stackables: Combining Tangibles for Faceted Browsing“. In: *Proceedings of the International Working Conference on Advanced Visual Interfaces*. AVI '12. New York, NY, USA: ACM, 2012, pp. 241–248 (cit. on pp. 36, 46).

[Kov+15] Adriana Kovashka, Devi Parikh, and Kristen Grauman. „WhittleSearch: Interactive Image Search with Relative Attribute Feedback“. In: *International Journal of Computer Vision* 115.2 (2015), pp. 185–210 (cit. on pp. 41, 42, 49).

[Koz+09] László Kozma, Arto Klami, and Samuel Kaski. „GaZIR: Gaze-based Zooming Interface for Image Retrieval“. In: *Proceedings of the 2009 International Conference on Multimodal Interfaces*. ICMI-MLMI '09. New York, NY, USA: ACM, 2009, pp. 305–312 (cit. on pp. 40, 41, 43).

[Kuh91] Carol C Kuhlthau. „Inside the search process: Information seeking from the user's perspective“. In: *Journal of the American Society for information Science* 42.5 (1991), p. 361 (cit. on p. 11).

[Lan+14] Ricardo Langner, Anton Augsburg, and Raimund Dachselt. „CubeQuery: Tangible Interface for Creating and Manipulating Database Queries“. In: *Proceedings of the Ninth ACM International Conference on Interactive Tabletops and Surfaces*. ITS ’14. New York, NY, USA: ACM, 2014, pp. 423–426 (cit. on pp. 36, 46).

[Lee+07] Gun A Lee, Gerard J Kim, and Mark Billinghurst. „Interaction design for tangible augmented reality applications“. In: *Emerging Technologies of Augmented Reality: Interfaces and Design*. IGI Global, 2007, pp. 261–282 (cit. on p. 19).

[Li+16] Qingyong Li, Mei Tian, Jun Liu, and Jinrui Sun. „An implicit relevance feedback method for CBIR with real-time eye tracking“. In: *Multimedia Tools and Applications* 75.5 (2016), pp. 2595–2611 (cit. on p. 43).

[LJ14] Fei Li and Hosagrahar V Jagadish. „NaLIR: An Interactive Natural Language Interface for Querying Relational Databases“. In: *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*. SIGMOD ’14. New York, NY, USA: ACM, 2014, pp. 709–712 (cit. on pp. 34, 44).

[LL14] Adam Lella and Andrew Lipsman. „The US mobile app report“. In: *Tech. Rep.* 8 (2014) (cit. on p. 1).

[Mas+97] A Massari, L Saladini, M Hemmje, and F Sisinni. „Virgilio: a non-immersive VR system to browse multimedia databases“. In: *Proceedings of IEEE International Conference on Multimedia Computing and Systems*. 1997, pp. 573–580 (cit. on pp. 38, 46, 47, 49).

[Mat+04] Krešimir Matković, Thomas Psik, Ina Wagner, and Werner Purgathofer. „Tangible Image Query“. In: *Smart Graphics: 4th International Symposium, SG 2004, Banff, Canada, May 23-25, 2004. Proceedings*. Ed. by Andreas Butz, Antonio Krüger, and Patrick Olivier. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 31–42 (cit. on pp. 35, 45).

[Nak+03] Munehiro Nakazato, Ljubomir Manola, and Thomas S Huang. „ImageGrouper: a group-oriented user interface for content-based image retrieval and digital image arrangement“. In: *Journal of Visual Languages & Computing* 14.4 (2003), pp. 363–386 (cit. on pp. 41, 42, 50).

[NH01] M Nakazato and T S Huang. „3D MARS: immersive virtual reality for content-based image retrieval“. In: *Multimedia and Expo, 2001. ICME 2001. IEEE International Conference on*. 2001, pp. 44–47 (cit. on pp. 39, 41, 42, 48).

[Nie+14] K van Nieuwenhuizen, G J Schaap, and W Hürst. „3D HMD Image Browsers: Optimal User Placement and Exploration Mechanisms“. Utrecht, 2014 (cit. on pp. 40, 47).

[Nor88] D A Norman. *The Psychology of Everyday Things*. Basic Books, 1988 (cit. on p. 10).

[NW08] G P Nguyen and M Worring. „Interactive Access to Large Image Collections Using Similarity-based Visualization“. In: *J. Vis. Lang. Comput.* 19.2 (2008), pp. 203–224 (cit. on pp. 14, 15).

[Pac+15] D Pacheco, S Wierenga, P Omedas, et al. „A location-based Augmented Reality system for the spatial interaction with historical datasets“. In: *2015 Digital Heritage*. Vol. 1. 2015, pp. 393–396 (cit. on p. 37).

[Pap+14] G T Papadopoulos, K C Apostolakis, and P Daras. „Gaze-Based Relevance Feedback for Realizing Region-Based Image Retrieval“. In: *IEEE Transactions on Multimedia* 16.2 (2014), pp. 440–454 (cit. on pp. 41, 43).

[PD15] Bernhard Preim and Raimund Dachselt. *Interaktive Systeme: Band 2: User Interface Engineering, 3D-Interaktion, Natural User Interfaces*. Springer-Verlag, 2015 (cit. on pp. 19–22).

[Qua+10] Novi Quadrianto, Kristian Kersting, Tinne Tuytelaars, and Wray L Buntine. „Beyond 2D-grids: A Dependence Maximization View on Image Browsing“. In: *Proceedings of the International Conference on Multimedia Information Retrieval*. MIR '10. New York, NY, USA: ACM, 2010, pp. 339–348 (cit. on p. 40).

[Rod+01] Kerry Rodden, Wojciech Basalaj, David Sinclair, and Kenneth Wood. „Does Organisation by Similarity Assist Image Browsing?“ In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '01. New York, NY, USA: ACM, 2001, pp. 190–197 (cit. on pp. 14, 15).

[Rui+98] Yong Rui, Thomas S Huang, Michael Ortega, and Sharad Mehrotra. „Relevance feedback: a power tool for interactive content-based image retrieval“. In: *IEEE Transactions on Circuits and Systems for Video Technology* 8.5 (1998), pp. 644–655 (cit. on p. 15).

[SA11] Klaus Schoeffmann and David Ahlstrom. „Similarity-Based Visualization for Image Browsing Revisited“. In: *Proceedings of the 2011 IEEE International Symposium on Multimedia*. ISM '11. Washington, DC, USA: IEEE Computer Society, 2011, pp. 422–427 (cit. on p. 15).

[Sch+13] Klaus Schoeffmann, David Ahlström, and Laszlo Böszörmenyi. „A User Study of Visual Search Performance with Interactive 2D and 3D Storyboards“. In: *Proceedings of the 9th International Conference on Adaptive Multimedia Retrieval: Large-scale Multimedia Retrieval and Evaluation*. AMR'11. Berlin, Heidelberg: Springer-Verlag, 2013, pp. 18–32 (cit. on p. 40).

[Sch+14] K Schoeffmann, D Ahlström, and M A Hudelist. „3-D Interfaces to Improve the Performance of Visual Known-Item Search“. In: *IEEE Transactions on Multimedia* 16.7 (2014), pp. 1942–1951 (cit. on p. 40).

[Sch10] Gerald Schaefer. „A Next Generation Browsing Environment for Large Image Repositories“. In: *Multimedia Tools Appl.* 47.1 (2010), pp. 105–120 (cit. on pp. 40, 47).

[Sch14] Klaus Schoeffmann. „The Stack-of-Rings Interface for Large-Scale Image Browsing on Mobile Touch Devices“. In: *Proceedings of the 22Nd ACM International Conference on Multimedia*. MM '14. New York, NY, USA: ACM, 2014, pp. 1097–1100 (cit. on pp. 40, 47).

[SG11] Grant Strong and Minglun Gong. „Similarity-based Image Organization and Browsing Using Multi-resolution Self-organizing Map“. In: *Image Vision Comput.* 29.11 (2011), pp. 774–786 (cit. on p. 15).

[SH16] Dieter Schmalstieg and Tobias Hollerer. *Augmented Reality: Principles and Practice*. Addison-Wesley Professional, 2016 (cit. on pp. 17–22, 28).

[Sha48] C E Shannon. „A mathematical theory of communication“. In: *The Bell System Technical Journal* 27.3 (1948), pp. 379–423 (cit. on p. 26).

[Sme+00] A W M Smeulders, M Worring, S Santini, A Gupta, and R Jain. „Content-based image retrieval at the end of the early years“. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22.12 (2000), pp. 1349–1380 (cit. on pp. 2, 8, 9, 12–14, 24–27, 57).

[Sna+06] Noah Snavely, Steven M Seitz, and Richard Szeliski. „Photo Tourism: Exploring Photo Collections in 3D“. In: *ACM Trans. Graph.* 25.3 (2006), pp. 835–846 (cit. on pp. 39, 41, 42).

[Sug+16] Hitoshi Sugimura, Hayato Tsukiji, Mizuki Kumada, Toshiya Iiba, and Kosuke Takano. „A Sketch-Based User Interface for Image Search Using Sample Photos“. In: *Human Interface and the Management of Information:Information, Design and Interaction: 18th International Conference, HCI International 2016 Toronto, Canada, July 17-22, 2016, Proceedings, Part I*. Ed. by Sakae Yamamoto. Cham: Springer International Publishing, 2016, pp. 361–370 (cit. on pp. 35, 45).

[TT00] G Y Tian and D Taylor. „Colour image retrieval using virtual reality“. In: *Information Visualization, 2000. Proceedings. IEEE International Conference on*. 2000, pp. 221–225 (cit. on pp. 39, 48).

[UJ06] Jana Urban and Joemon M Jose. „Can a Workspace Help to Overcome the Query Formulation Problem in Image Retrieval?“ In: *Proceedings of the 28th European Conference on Advances in Information Retrieval*. ECIR'06. Berlin, Heidelberg: Springer-Verlag, 2006, pp. 385–396 (cit. on pp. 15, 16).

[UJ07] Jana Urban and Joemon M Jose. „Evaluating a Workspace's Usefulness for Image Retrieval“. In: *Multimedia Syst.* 12.4-5 (2007), pp. 355–373 (cit. on p. 16).

[Ull+03] Brygg Ullmer, Hiroshi Ishii, and Robert J K Jacob. „Tangible query interfaces: Physically constrained tokens for manipulating database queries“. In: *Proc. of INTERACT*. Vol. 3. 2003, pp. 279–286 (cit. on pp. 36, 46).

[WF09] Sean White and Steven Feiner. „SiteLens: Situated Visualization Techniques for Urban Site Visits“. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '09. New York, NY, USA: ACM, 2009, pp. 1117–1120 (cit. on pp. 18, 26, 28).

[Wor+07] Marcel Worring, Ork de Rooij, and Ton van Rijn. „Browsing Visual Collections Using Graphs“. In: *Proceedings of the International Workshop on Workshop on Multimedia Information Retrieval*. MIR '07. New York, NY, USA: ACM, 2007, pp. 307–312 (cit. on p. 14).

[Yee+03] Ka-Ping Yee, Kirsten Swearingen, Kevin Li, and Marti Hearst. „Faceted Metadata for Image Search and Browsing“. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '03. New York, NY, USA: ACM, 2003, pp. 401–408 (cit. on pp. 36, 46).

[Yeh+12] Kai-Chen Yeh, Meng-Han Tsai, and Shih-Chung Kang. „On-Site Building Information Retrieval by Using Projection-Based Augmented Reality.“ In: *Journal of Computing in Civil Engineering* 26.3 (2012), pp. 342–355 (cit. on p. 37).

[Zha+14] Qi Zhang, Simon Zaaijer, Song Wu, and Michael S Lew. „3D Image Browsing: The Planets“. In: *Proceedings of International Conference on Multimedia Retrieval*. ICMR '14. New York, NY, USA: ACM, 2014, 511:511–511:513 (cit. on p. 38).

[Zwo+10] Roelof van Zwol, Börkur Sigurbjornsson, Ramu Adapala, et al. „Faceted Exploration of Image Search Results“. In: *Proceedings of the 19th International Conference on World Wide Web*. WWW ’10. New York, NY, USA: ACM, 2010, pp. 961–970 (cit. on pp. 41, 42).

Glossary

API application programming interface. 64, 67, 69, 70, 81, 90
AR augmented reality. 1, 2, 3, 4, 5, 16, 17, 18, 19, 20, 21, 22, 23, 25, 27, 28, 29, 33, 36, 38, 42, 43, 44, 45, 46, 47, 48, 49, 50, 52, 55, 57, 60, 63, 75, 78, 80, 91, 93, 107, 109

CBIR content-based image retrieval. 2, 5, 13

GUI graphical user interface. 36

HCIR human computer information retrieval. 2, 13
HMD head-mounted display. 1, 2, 3, 5, 16, 20, 40, 60, 75, 78, 80
HTP HoloLens Technical Preview. 71

IDE integrated development environment. 71
ISO International Organization for Standardization. 79

JSON JavaScript Object Notation. 64, 69

MVC model view controller. 67

REST representational state transfer. 67, 69
ROI region of interest. 15

SUS System Usability Scale. 80, 87

TUI tangible user interface. 36

UI user interface. 21, 65, 66, 67, 68, 72
USB Universal Serial Bus. 72

VR virtual reality. 1, 36, 38, 39

WLAN wireless local area network. 72

List of Figures

List of Tables

Colophon

This thesis was typeset with LaTeX 2_ε. It uses the *Clean Thesis* style developed by Ricardo Langner. The design of the *Clean Thesis* style is inspired by user guide documents from Apple Inc.

CleanThesis

Download the *Clean Thesis* style at `http://cleanthesis.der-ric.de/`.

Eigenständigkeitserklärung

Hiermit versichere ich, dass ich die vorliegende Arbeit mit dem Titel *Image Retrieval within Augmented Reality* selbstständig und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Alle Stellen, die anderen Quellen im Wortlaut oder dem Sinn nach entnommen wurden, sind als solche mit Angaben der Herkunft kenntlich gemacht. Dies gilt auch für alle bildlichen Darstellungen. Die Arbeit wurde bisher in gleicher oder ähnlicher Form noch nicht als Prüfungsleistung eingereicht.

Dresden, 05. Mai 2017

Philip Manja